BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-051889
(43)Date of publication of application : 20.02.1998

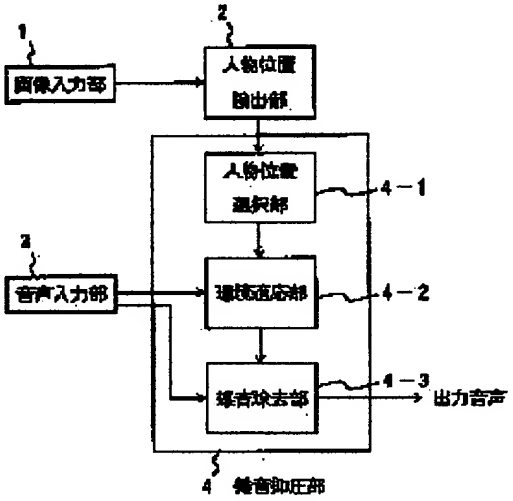
(51)Int.Cl. H04R 3/00
H03H 17/00
H03H 21/00

(21)Application number : 08-206210
(22)Date of filing : 05.08.1996

(71)Applicant : TOSHIBA CORP
(72)Inventor : NAGATA HITOSHI

(54) DEVICE AND METHOD FOR GATHERING SOUND

(57)Abstract:
PROBLEM TO BE SOLVED: To extract all sounds from person's positions or only the sound from a specific person's position by determining a filter coefficient according to a person's position to be processed which is selected out of detected person's positions or a signal regarding the specific position and a sound source signal.
SOLUTION: A person's position detection part 2 detects the position of a person from an image inputted from an image input part 1. A noise suppression part 4 selects the person's position to be processed out of detected person's positions by a person's position selection part 4-1 and generates the learning signal of an adaptive filter according to the person's position selected by an environment adaptation part 4-2 to determine the coefficient of the adaptive filter. A noise removal part 4-3 filters an input sound from a sound input part 3 by using the determined filter coefficient to suppress its noise. Consequently, when plural persons speak at the same time, all the sounds of the persons can be extracted while having their background noise suppressed or the sound of one person can be extracted while the sounds of other persons are suppressed.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－51889

(43)公開日　平成10年(1998)２月20日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｈ０４Ｒ	3/00	３２０		Ｈ０４Ｒ	3/00	３２０
Ｈ０３Ｈ	17/00	６０１	9274－5Ｊ	Ｈ０３Ｈ	17/00	６０１Ｇ
	21/00		9274－5Ｊ		21/00	

審査請求　未請求　請求項の数10　ＯＬ　（全 26 頁）

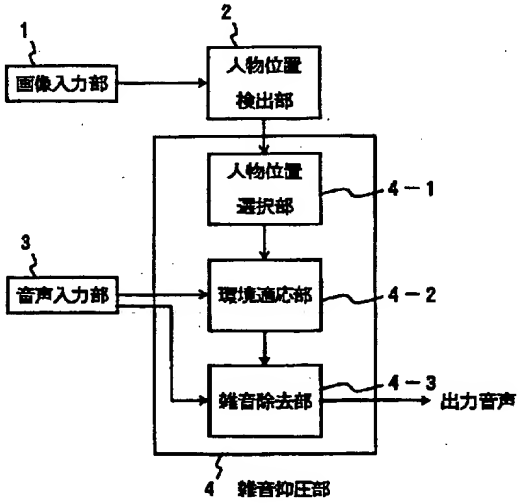
(21)出願番号　特願平8－206210

(22)出願日　平成８年(1996)８月５日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　永田　仁史
大阪府大阪市北区大淀中１丁目１番30号
株式会社東芝関西支社内
(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】　音声収集装置及び音声収集方法

(57)【要約】
【課題】複数の人物位置からの音声に対して、背景雑音を抑えてすべての音声を同時に抽出するがあるいは、特定の人物位置からの音声のみを抽出する。
【解決手段】入力された画像情報を処理して複数の人物位置を求める人物位置検出部２と、検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択部４－１と、任意に発生された音源信号を、人物位置選択部４－１によって選択された人物位置に配置したものとしたときに観測して得られる入力信号と、選択された人物位置からのすべての音声に対する感度を同時に高くするモードと、選択された人物位置のうち、特定の目的位置からの音声のみを高くするモードのうちいずれかの選択に応じて音源信号から生成される応答信号とに基づいて、フィルタ係数を決定する環境適応部４－２と、決定されたフィルタ係数を用いて、入力された所望の人物位置からの音声のみを抽出する雑音除去部４－３とを具備する。



1

【特許請求の範囲】

【請求項１】　複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、
複数のチャネルを介して個々に音声を入力する音声入力手段と、
前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、
この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、
任意に生成した音源信号を、前記人物位置選択手段によって選択された人物位置に配置したものとしたときに観測して得られる第１の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第２の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定手段と、
このフィルタ係数決定手段によって決定されたフィルタ係数を用いて、前記音声入力手段によって入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出手段と、
を具備することを特徴とする音声収集装置。

【請求項２】　前記選択された人物位置のうち、前記特定の目的位置からの音声のみを高くするモードにおいて、複数の目的位置に対応して前記フィルタ係数決定手段及び音声抽出手段を複数個設け、複数の人物位置からの音声を分離して抽出するようにしたことを特徴とする請求項１記載の音声収集装置。

【請求項３】　テスト発声データの入力と前記音声入力手段を介して入力される通常の音声入力の切り替えを指示する入力モード切り替え手段と、入力モードがテスト発声データ入力であるときに、取り込んだテスト発声データのレベルを求めるテスト発声レベル計算手段とをさらに具備することを特徴とする請求項１または２記載の音声収集装置。

【請求項４】　前記画像入力手段によって入力された画像から人物の発声動作に関する情報を位置別に検出する位置別発声動作情報検出手段をさらに具備し、前記フィルタ係数決定手段は、検出した位置別の発声動作に関する情報と、入力された音声から求めた位置別到来パワーの少なくとも一方に基づいて、前記第１の信号である入力信号と前記第２の信号である希望応答信号とを生成することを特徴とする請求項１乃至３のいずれかに記載の音声収集装置。

【請求項５】　複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、
複数のチャネルを介して個々に音声を入力する音声入力手段と、

2

前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、
この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、
この人物位置選択手段によって選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定手段と、
このフィルタ制約設定手段の制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力手段によって入力される音声にフィルタ処理を施して音声を抽出する音声抽出手段と、
を具備することを特徴とする音声収集装置。

【請求項６】　前記フィルタ制約設定手段は、前記選択された人物位置の数が複数の場合に、この複数の人物位置の中の一つの位置を目的位置とし、該目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする第１の制約と、前記目的位置以外の人物位置からの音声に対しては、選択されなかった人物位置と比較して感度を低くする第２の制約をフィルタ処理の制約として設定し、前記音声抽出手段は、前記第１、第２の制約の基にフィルタ出力を最小化してフィルタ係数を決定することを特徴とする請求項５記載の音声収集装置。

【請求項７】　複数の人物を撮影して得られた画像を入力する画像入力手段と、
この画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、
この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、
複数のチャネルを介して個々に音声を入力する音声入力手段と、
前記人物位置選択手段によって選択された少なくとも一つの人物位置の中の一つの位置を目的位置とし、この目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする制約を設定するフィルタ制約設定手段と、
任意に作成した音源信号を、前記目的位置以外の人物位置に配置したものとしたときに観測される信号を生成する入力信号生成手段と、
前記制約のもとで前記入力信号に基づき目的位置以外の人物位置からの音声に対して感度を低くするようにフィルタを決定するフィルタ決定手段と、
このフィルタ決定手段によって求められたフィルタ係数を用いて、前記音声入力手段によって入力された音声にフィルタ処理を施して音声を抽出する音声抽出手段と、
を具備することを特徴とする音声収集装置。

【請求項８】　前記フィルタ制約設定手段は、前記選択された人物位置の中から複数の目的位置を設定した場合に、この複数の目的位置の一つからの音声に対する感度を、選択されなかった人物位置と比較して高くする制約をフィルタ処理の制約として設定し、前記目的位置以外の人物位置に音源があるものとしたときに観測される入力信号に基づき、前記目的位置以外の人物位置からの音声に対しては感度を、選択されなかった人物位置と比較して低くするようにフィルタを設定するフィルタ設定手段と音声抽出手段とを、前記目的位置の変更に対応して複数個設け、複数の人物位置からの音声を分離して抽出するようにしたことを特徴とする請求項７記載の音声収集装置。

【請求項９】　複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、
複数のチャネルを介して個々に音声を入力する音声入力工程と、
前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、
この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、
任意に生成した音源信号を、前記人物位置選択工程で選択された人物位置に配置したものとしたときに観測して得られる第１の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第２の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定工程と、
このフィルタ係数決定工程において決定されたフィルタ係数を用いて、前記音声入力工程において入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出工程と、
を具備することを特徴とする音声収集方法。

【請求項１０】　複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、
複数のチャネルを介して個々に音声を入力する音声入力工程と、
前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、
この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、
この人物位置選択工程において選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定工程と、
このフィルタ制約設定工程における制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力工程において入力される音声にフィルタ処理を施して音声を抽出する音声抽出工程と、
を具備することを特徴とする音声収集方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は音声収集装置及び音声収集方法に関し、特に、音声認識装置やテレビ会議システムなどにおいて、雑音を取り除いて目的の音声を取り出す雑音抑圧技術に関する。

【０００２】

【従来の技術】音声認識装置やテレビ会議システムにおいては、高品質な音声の収録のため、指向性マイクロホンやマイクロホンアレイを使った雑音抑圧技術が提案されている。特に、テレビ会議システムの分野では、複数の会議参加者の中から発言者の音声と画像を自動的に得るため、ビデオカメラの画像を処理して得られた移動物体の位置に基づいて複数のマイクロホンの信号を処理する方法が例えば特開平５－２２７５３１号公報に開示されている。

【０００３】しかしながら、この方法ではマイクロホンアレイからの信号を、一つの目的の人物位置からの音声に対して位相が一致するようにした遅延和法により処理しているため、他の方向から到来した雑音に対する抑圧性能は十分でないという問題があった。

【０００４】一方、マイクロホンアレイの出力を処理して効果的に雑音を抑圧する技術としては、適応フィルタを指向性制御に用いた適応マイクロホンアレイ技術が従来より知られており、例えば文献（電子情報通信学会編　音響システムとデジタル処理　ｐｐ．１７１－２１８）に詳述されている。適応マイクロホンアレイ処理では、雑音の到来方向を知る必要はないが、目的とする音波の到来方向は既知として処理するのが一般的である。音波の到来方向は、マイクロホンアレイからの信号を処理して推定することもできるが、発声中のみしか検出できないため処理の安定性に問題がある。

【０００５】これに対し、画像を処理して得られた人物位置を目的音の到来方向として用いる方法が知られており、この場合は、発声していないときにも位置が推定できるため安定であり、例えば文献（ＩＣＡＳＳＰ’９５「Knowing Who to Listen toin Special Recognition Visually Guided Beamforming」ｐｐ８４８－８５１）に開示されている。

【０００６】

【発明が解決しようとする課題】しかしながら、上記文献をはじめとする従来の開示技術においては、画像の処理により人物位置が複数検出された場合に対する対処方法がないため、目的としない人物から発声があった場合はそれを除くような適応処理を行っていた。ところが、

この適応処理が完了するまでに妨害音が混入してしまったり、複数の話者が同時に発声した場合に、注目している一人の音声以外はクリアに入力できない、という問題があった。

【０００７】本発明はこのような課題に着目してなされたものであり、その目的とするところは、複数の人物位置からの音声に対して、背景雑音を抑えてすべての音声を同時に抽出するかあるいは、特定の人物位置からの音声のみを抽出することが可能な音声収集装置及び音声収集方法を提供することにある。

【０００８】

【課題を解決するための手段】上記の目的を達成するために、第１の発明に係る音声収集装置は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、複数のチャネルを介して個々に音声を入力する音声入力手段と、前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、任意に生成した音源信号を、前記人物位置選択手段によって選択された人物位置に配置したものとしたときに観測して得られる第１の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第２の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定手段と、このフィルタ係数決定手段によって決定されたフィルタ係数を用いて、前記音声入力手段によって入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出手段とを具備する。

【０００９】また、第２の発明に係る音声収集装置は、第１の発明において、前記選択された人物位置のうち、前記特定の目的位置からの音声のみを高くするモードにおいて、複数の目的位置に対応して前記フィルタ係数決定手段及び音声抽出手段を複数個設け、複数の人物位置からの音声を分離して抽出する。

【００１０】また、第３の発明に係る音声収集装置は、第１または第２の発明において、テスト発声データの入力と前記音声入力手段を介して入力される通常の音声入力の切り替えを指示する入力モード切り替え手段と、入力モードがテスト発声データ入力であるときに、取り込んだテスト発声データのレベルを求めるテスト発声レベル計算手段とをさらに具備する。

【００１１】また、第４の発明に係る音声収集装置は、第１乃至第３の発明のいずれかにおいて、前記画像入力手段によって入力された画像から人物の発声動作に関する情報を位置別に検出する位置別発声動作情報検出手段をさらに具備し、前記フィルタ係数決定手段は、検出した位置別の発声動作に関する情報と、入力された音声から求めた位置別到来パワーの少なくとも一方に基づいて、前記第１の信号である入力信号と前記第２の信号である希望応答信号とを生成する。

【００１２】また、第５の発明の係る音声収集装置は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力手段と、複数のチャネルを介して個々に音声を入力する音声入力手段と、前記画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、この人物位置選択手段によって選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定手段と、このフィルタ制約設定手段の制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力手段によって入力される音声にフィルタ処理を施して音声を抽出する音声抽出手段とを具備する。

【００１３】また、第６の発明に係る音声収集装置は、第５の発明において、前記フィルタ制約設定手段は、前記選択された人物位置の数が複数の場合に、この複数の人物位置の中の一つの位置を目的位置とし、該目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする第１の制約と、前記目的位置以外の人物位置からの音声に対しては、選択されなかった人物位置と比較して感度を低くする第２の制約をフィルタ処理の制約として設定し、前記音声抽出手段は、前記第１、第２の制約の基にフィルタ出力を最小化してフィルタ係数を決定する。

【００１４】また、第７の発明に係る音声収集装置は、複数の人物を撮影して得られた画像を入力する画像入力手段と、この画像入力手段によって入力された画像情報を処理して複数の人物位置を求める人物位置検出手段と、この人物位置検出手段によって検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択手段と、複数のチャネルを介して個々に音声を入力する音声入力手段と、前記人物位置選択手段によって選択された少なくとも一つの人物位置の中の一つの位置を目的位置とし、この目的位置からの音声に対する感度を、選択されなかった人物位置と比較して高くする制約を設定するフィルタ制約設定手段と、任意に作成した音源信号を、前記目的位置以外の人物位置に配置したものとしたときに観測される信号を生成する入力信号生成手段と、前記制約のもとで前記入力信号に基づき目的位置以外の人物位置からの音声に対して感度を低くするようにフィルタを決定するフィルタ決定手段と、このフィルタ決定手段によって求められたフィルタ

係数を用いて、前記音声入力手段によって入力された音声にフィルタ処理を施して音声を抽出する音声抽出手段とを具備する。

【００１５】また、第８の発明に係る音声収集装置は、第７の発明において、前記フィルタ制約設定手段は、前記選択された人物位置の中から複数の目的位置を設定した場合に、この複数の目的位置の一つからの音声に対する感度を、選択されなかった人物位置と比較して高くする制約をフィルタ処理の制約として設定し、前記目的位置以外の人物位置に音源があるものとしたときに観測される入力信号に基づき、前記目的位置以外の人物位置からの音声に対しては感度を、選択されなかった人物位置と比較して低くするようにフィルタを設定するフィルタ設定手段と音声抽出手段とを、前記目的位置の変更に対応して複数個設け、複数の人物位置からの音声を分離して抽出する。

【００１６】また、第９の発明に係る音声収集方法は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、複数のチャネルを介して個々に音声を入力する音声入力工程と、前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、任意に生成した音源信号を、前記人物位置選択工程で選択された人物位置に配置したものとしたときに観測して得られる第１の信号と、前記選択された人物位置からのすべての音声に対する感度を、選択されなかった人物位置と比較して同時に高くするモードと、前記選択された人物位置のうち、特定の目的位置からの音声のみを、選択されなかった人物位置と比較して高くするモードのうちいずれかの選択に応じて前記音源信号から生成される第２の信号とに基づいて、フィルタ係数を決定するフィルタ係数決定工程と、このフィルタ係数決定工程において決定されたフィルタ係数を用いて、前記音声入力工程において入力された音声のうち、前記選択されたモードに対応する音声のみを抽出する音声抽出工程とを具備する。

【００１７】また、第１０の発明に係る音声収集方法は、複数の人物の少なくとも一部を撮影して得られた画像を入力する画像入力工程と、複数のチャネルを介して個々に音声を入力する音声入力工程と、前記画像入力工程において入力された画像情報を処理して複数の人物位置を求める人物位置検出工程と、この人物位置検出工程において検出された複数の人物位置から、処理対象となる人物位置を少なくとも一人選択する人物位置選択工程と、この人物位置選択工程において選択された人物位置に基づいて、前記少なくとも一人の人物位置からの音声に対する感度を同時に一定の値にする制約をフィルタ処理の制約として設定するフィルタ制約設定工程と、このフィルタ制約設定工程における制約に基づいてフィルタ係数を決定し、このフィルタ係数を用いて前記音声入力工程において入力される音声にフィルタ処理を施して音声を抽出する音声抽出工程とを具備する。

【００１８】

【発明の実施の形態】まず、本実施形態の概略を説明する。本実施形態では画像から人物位置を検出し、その人物位置に基づいて適応マイクロホンアレイにより雑音を抑圧して音声を抽出するシステムにおいて、特に、複数の人物位置が検出された場合に対処するため、以下の方法を用いる。

【００１９】すなわち、第１の概略においては、雑音を抑圧するフィルタの決定に適応フィルタの処理方式の一つであるパイロット信号法を利用し、画像の処理により得られた複数の人物の位置に基づき適応フィルタの学習信号である入力信号と希望応答信号を人工的に生成し、複数の人物位置から到来する音波に対して同時に感度を一定に保つように制御を行なうことにより背景雑音を抑えてすべての話者の音声を同時に取り出すことを可能にするものである。

【００２０】また、同じ構成により、目的の人物位置から到来する音波については感度が高く、他の人物位置から到来する音波については感度が低くなるように制御を行なうことにより、特定の人物のみの音声を取り出すことも可能にしている。パイロット信号法に関しては、上記文献（音響システムとデジタル処理）または文献（ＰＲＯＣ．ＩＥＥＥ Ｖｏｌ．５５，Ｎｏ１２（１９６７）、B Widraw:「Adaptive Antenna systems」に詳述されている。

【００２１】また、第２の概略においては、適応フィルタによる雑音抑圧のフィルタ係数の学習の際、人物位置から到来する音波のパワーを推定し、このパワーに基づいて適応フィルタの入力信号の振幅と適応フィルタの収束速度を決定しているため、人工的に生成する信号を実際の環境に忠実に合わせることができ、精度良く雑音抑圧のフィルタを学習することができる。

【００２２】また、第３の概略においては、適応フィルタによる雑音抑圧のフィルタ係数の学習の際、人物位置から到来する音波のパワーに加え、人物の画像から発声動作を表す情報を抽出しており、これらの位置ごとのパワーと位置ごとの発声動作に関する情報から、適応フィルタの入力信号の振幅と適応フィルタの収束速度を決定しているため、背景雑音が大きく、到来パワーの推定精度が低い場合でも精度良く雑音抑圧のフィルタを学習することができる。

【００２３】また、第４の概略においては、テスト発声を収集するための入力モード切り替え手段を具備することにより、テスト発声データを入力し、その入力レベルからパイロット信号法によって適応フィルタ処理を行う際の学習信号の振幅を正確に決定して適応フィルタの学

習を行うことにより、高精度な雑音抑圧を可能としている。

【００２４】また、第５の概略においては、目的の人物から到来する音に対しては感度を一定に保ち、他の人物から到来する音に対しては感度を低くするように適応フィルタの学習信号を生成してフィルタ係数を学習し、人物ごとにこのようなフィルタを用意することにより、複数の人物が発声した場合に、各人物ごとの発声を別個に取り出すことを可能にしている。

【００２５】また、第６の概略においては、パイロット信号法による適応フィルタ処理の代わりに、拘束条件付き適応フィルタ処理を用い、画像の処理により得られた複数の人物の位置に対してマイクロホンアレイの感度を一定に保つという拘束条件のもとで適応フィルタの係数を決定することにより、背景雑音を抑えてすべての話者の音声を同時に取り出すことを可能にしている。この方式では、パイロット信号法で必要だった発声区間に応じた適応の制御が不要であり、より少ない構成要素で実現できる。

【００２６】同概略においては、目的の人物から到来する音に対しては感度を一定に保ち、他の人物から到来する音に対しては感度を低くするという制約のもとで雑音抑圧のフィルタを決定することにより、特定の人物の音声だけを取り出すことも可能である。

【００２７】また、第７の概略においては、第６の概略において用いた拘束条件付き適応フィルタ処理を用い、目的の人物から到来する音に対しては感度を一定に保ち、他の人物から到来する音に対しては感度を低くするという制約のもとで雑音抑圧のフィルタの係数を決定し、人物ごとにこのようなフィルタを用意することにより、複数の人物が発声した場合に、各人物ごとの発声を別個に取り出すことを可能にしている。

【００２８】また、第８の概略においては、第６、第７の概略で用いた拘束条件付き適応フィルタ処理と第１から第５の概略で用いたパイロット信号法を組み合わせ、目的の人物から到来する音に対しては感度を一定に保ち、制約のもとで雑音抑圧のフィルタの係数を決定し、他の人物位置から到来する音に対しては感度が低くなるように学習信号を生成し、この学習信号により拘束条件付き適応フィルタによりフィルタ係数を決定することにより、拘束条件が多くなることによる雑音抑圧性能の低下を抑え、また学習信号生成の演算量を減らして同等の性能を実現している。

【００２９】また、第９の概略においては、第８の概略において雑音抑圧のフィルタを人物ごとに複数用意することにより、複数の人物が発声した場合に、各人物ごとの発声を別個に取り出すことを可能にしている。

【００３０】以下に図面を参照して上記した実施形態を詳細に説明する。

【００３１】まず、図１を参照しながら、第１実施形態について説明する。本実施形態は、画像を処理して検出した人物位置に基づいて適応フィルタの学習信号を生成し、学習したフィルタにより雑音抑圧処理を行うものである。本実施形態では、複数の人物位置を対象として適応フィルタ学習の制御を行えるようにしているため、従来１個の対象にしか考えられてこなかった雑音抑圧処理を、複数の対象に対して行え、会話や複数話者の同時発声の場合の音声入力を自動的に安定して高対雑音比で行うことが可能である。

【００３２】図１において、１はビデオカメラなどから画像を入力する画像入力部、２は入力した画像を処理して人物の位置を検出する人物位置検出部、３は複数のマイクロホンからの音声を並列に入力する音声入力部、４は画像処理により検出された人物位置に基づいて複数のマイクロホンからの入力音声から雑音を抑圧して音声を取り出す雑音抑圧部であり、その内部構成は、人物位置検出部２により検出された人物位置の中から処理対象とする人物位置を選択する人物位置選択部４－１と、選択された人物位置に基づいて適応フィルタの学習を行う環境適応部（フィルタ係数決定手段）４－２と、決定されたフィルタ係数により雑音抑圧処理を行う雑音除去部４－３とからなる。

【００３３】この構成において、画像入力部１より入力した画像を人物位置検出部２に送り、人物位置検出部２において人物の方向または位置を画像から検出する。検出した人物位置の中から処理対象とする人物位置を人物位置選択部４－１において選択し、環境適応部４－２において、前記選択された人物位置に基づいて適応フィルタの学習信号を生成して適応フィルタの係数を決定し、雑音除去部４－３において決定したフィルタ係数を用いて入力音声にフィルタ処理を行って雑音を抑圧する。

【００３４】以下に上記した処理をさらに詳細に説明する。まず、画像からの人物位置の検出について説明する。画像からの人物位置の検出は、テンプレートマッチングに基づいた顔領域の抽出処理などにより行なうことができ、テンプレートマッチングについては例えば「画像解析ハンドブック」（東京大学出版会）に詳述されている。さらに、テンプレートマッチングを用いて画像中の物体の動きを追跡することができ、例えば情報処理学会技術報告ＣＶ７６－７，ｐｐ．４９－５６（１９９２），小杉他：「シーンの中の顔の探索と認識」に詳述されている。これらの開示技術により、同じ人物の座標を時間ごとに特定し追跡できることが知られている。なお、これらの技術では、人物の顔を含む小領域の画像を扱うため、人物位置の検出とともにこれらの画像も得ることができる。

【００３５】画像による人物位置の検出では、一般に、画像入力に１個のビデオカメラを用いた場合、人物のカメラに対する方向は十分な精度で特定することができるが、カメラと人物の間の距離方向の測定は誤差が大き

い。それでも、人物の大きさの比較により、距離方向に関しておおまかな相対関係を得ることは可能である。ステレオカメラを用いた場合は、距離に関しても高精度に測定できるが、本実施形態では、人物の方向とおおまかな距離関係がわかれば十分であるため、カメラ１個による人物位置の抽出手法を用いればよく、ステレオカメラは必ずしも必要ない。もちろんステレオカメラあるいは多数のカメラを用いても差し支えない。人物位置の検出方法は本実施形態の本質に関わりがないので詳しい説明は省略するが、現状で利用可能な技術であることは言うまでもない。

【００３６】ビデオカメラとマイクロホンを組み合わせて処理を行う際、ビデオカメラとマイクロホンの位置の設定は種々考えられるが、ビデオカメラが１個の場合、例えば図２（ａ）に示すように設定する。ビデオカメラ５とマイクロホンアレイ６は人物から見て同じ方向にあるように設置し、マイクロホンアレイ６の処理とビデオカメラ５とで共通の方位座標を使うようにするのが望ましい。

【００３７】なお、図２（ｂ）に示すように、ビデオカメラ５を複数使う場合は、人物位置が３次元座標として得られるので、先のようにカメラ位置とマイクロホン位置を利用者から見て同じ方向に設定する必要はなく、マイクロホンアレイ処理の際、ビデオカメラ５から得られた人物座標をマイクロホンアレイ６からみた角度に変換して用いることもできる。

【００３８】以上の処理により人物位置が得られた後、雑音抑圧部４においては、人物位置選択部４－１により人物位置の中から処理対象とする人物位置を予め決めた数だけ選択し、該選択した人物位置に基づき、環境適応部４－２により適応フィルタの学習信号を生成して適応フィルタに入力し、フィルタ係数を決定する。そして、決定したフィルタ係数を用い、雑音除去部４－３で複数のマイクロホンからの入力音声に上記のフィルタによるフィルタ処理を行って出力音声を取り出す。

【００３９】複数のマイクロホンから入力した音声を処理して雑音を抑圧するための適応フィルタとしては種々のものが知られており、例えば文献（Haykin著：Adaptive Filter Theory）に詳述されているが、本実施形態では、複数の任意の方向あるいは位置から到来する音に対するアレイの応答を比較的簡単に設定できるパイロット信号法を用いている。

【００４０】以下に、雑音抑圧部４の詳細を説明する。まず、雑音抑圧部４では、人物位置選択部４－１において複数の人物位置の中から、音声の抽出処理を行う対象の人物位置を選択する。この選択においては、選択する人物位置の最大数をＮ、例えばＮ＝３とし、人物位置検出部２で特定された人物位置の数がＮより大きい場合に、特定された人物位置の中からＮ個の位置を選択し、小さい場合はすべてを選択する。選択の方法に関しては、例えば、カメラと人物の距離を基準とし、この距離が小さい順にＮ個の位置を用いるようにしてもよいし、カメラの中心方向と人物方向の角度差を基準とし、この角度差が小さい順にＮ個の位置を用いてもよい。

【００４１】また、上記２つの選択基準を組み合わせた値を基準としてもよい。すでに述べたように、使用するカメラが１個の場合で、カメラと人物の間の距離の値を得ることが困難な場合は、人物の大きさまたは人物の顔の大きさを人物位置とカメラとの距離の目安として使うことが可能である。

【００４２】例えば、図３に示すような画像データから図４に示すような人物位置の方向（Ｘ，Ｙ）と顔の大きさ（Ａ）、および人物方向から計算されるカメラ中心線方向と人物方向との角度差（Ｂ）が得られた場合、顔部分の面積が大きいほどカメラに近いとしてこの面積が大きい順に人物の番号６，４，３の３人を選択してもよいし、カメラ中心線方向と人物方向との角度差が小さい順に人物番号４，２，５を選択してもよいし、上記Ａ，Ｂを組み合わせた値、例えばＡ／Ｂの値の大きい順に人物番号４，３，６を選択してもよい。

【００４３】次に、パイロット信号法による適応フィルタ処理を行うため、環境適応部４－２は、図５に示すような構成としている。図５において、４－２ａは入力信号生成部、４－２ｄは希望応答生成部、４－２ｅは適応処理部、４－２ｃは学習信号レベル計算部、４－２ｂは音源信号生成部である。

【００４４】この構成において、まず、音源信号生成部４－２ｂにより、選択された人物位置ごとに音源があるものと仮定してその発生信号を生成し、学習信号レベル計算部４－２ｃにより、入力音声に基づいて入力信号生成の際の音源信号のレベルを決定する。次に、求められた学習信号レベルと音源信号とから、入力信号生成部４－２ａにより、選択された人物位置に基づき、適応フィルタの入力信号を生成すると同時に、学習信号レベルと音源信号とから、希望応答生成部４－２ｄにより適応フィルタの希望応答を生成し、生成した入力信号と希望応答を適応処理部４－２ｅに入力し、適応フィルタの適応処理を行う。適応フィルタの処理方式は、よく知られたＬＭＳでもまた、ＲＬＳでもよく、文献（ヘイキン著：適応フィルタ入門）に詳述されている。ここでは、ＬＭＳ適応フィルタにより説明する。

【００４５】適応フィルタの処理は、複数チャネルの入力各々に対し、図６に示すような遅延線タップ付きフィルタから構成されるユニバーサル型フィルタを用いて行うようにしている。図６において、フィルタのタップ数をＪ、ｉ番目のマイクロホンのフィルタ係数をw_{ij}，（$1 \leq i \leq N$，$1 \leq j \leq J$）としており、Ｊは例えば２００を用いる。この構成において、ｉ番目のマイクロホンの波形を$x_i(n)$とし、時刻ｎにおいてＪサンプル過去から時刻ｎまでの各マイクロホンの波形サンプルの系

列x_i ＝（$x_{i(n-J+1)}$），$x_{i(n-J+2)}$，…，$x_{i(n-1)}$，＊　＊$x_{i(n)}$）を全マイクロホンについて並べ、

$$X=(x_1, x_2, \cdots, x_N)^T \quad (1)$$

と、ベクトルで表す。また、ｉ番目のマイクロホンのフ※　※ィルタ係数w_{ij}を並べてベクトルで表して

$$w_i = (w_{i1}, w_{i2}, \cdots, w_{iJ}) \quad (2)$$

とし、さらに全マイクロホンについて並べて

$$W=(w_1, w_2, \cdots, w_J)^T \quad (3)$$

と表す。式（１）、（３）から、フィルタの出力は、

$$Y=W^H X \quad (4)$$

と表される。ここでフィルタ係数Ｗの要素は複素数とし、Ｈはベクトルの複素共役転置を表すものとする。Ｘは一般にスナップショットと呼ばれる。

【００４６】ＬＭＳ適応フィルタ（Normalized　LM　☆　☆Ｓ）による適応処理部４－２ｅでは、上記のフィルタ構造において次式に従ってフィルタ係数を更新し、フィルタ係数の学習を行う。

【００４７】

$$W_j = W_{j-1} - a * e * X / 2p \quad (5)$$

ここで、W_j はｊ回の更新後のフィルタ係数、ｅは誤差信号ｅ＝ｄ－W^H Ｘ、ｄは希望応答、ｐは希望応答のパワー、ａはステップサイズであり、０＜ａ＜１．０の範囲で実験的に決められるが、例えば０．１などが用いられる。

【００４８】上記のフィルタ更新に用いる入力信号Ｘと希望応答ｄは、人物位置に基づき入力信号生成部４－２ａと希望応答生成部４－２ｄで音源信号から各々生成する。これらの信号は人工的に生成するものであり、信号の内容によって雑音抑圧の仕方を制御することができる。例えば、選択された人物位置すべてから到来する音波に対して感度を高くする（Ａ）ことや、選択された人物位置のうち、ある人物位置からの音波に対しては感度を高くするが、それ以外に対しては抑圧する（Ｂ）などのように制御できる。

【００４９】今後、上記２つの抑圧処理の仕方を、抑圧処理のモード（Ａ），（Ｂ）と呼ぶことにする。特にモード（Ｂ）は、妨害音の発生する可能性の大きい方向に対して事前に感度を低くする方法であり、従来の適応マイクロホン処理で行われていたように、妨害音が発生してからその環境に適応して抑圧する手法よりも大幅に高品質な音声入力が行える。抑圧処理モードの設定は、初期設定の際に環境適応部４－２において設定するようにする。

【００５０】フィルタ更新に用いる入力信号Ｘと希望応答ｄの生成の前段階として、まず、音源信号生成部４－２ｂにおいて、人物位置の数の信号系列である音源信号を発生する。発生した音源信号の内容は人工的なものでよく、例えば、ランダム雑音でもよい。このとき、ランダム雑音は人物位置ごとに無相関となるようにするため、人物位置ごとに独立な乱数系列から生成するようにする。また、周波数特性は、例えば平均的な音声のスペクトルの傾きと同じになるようにフィルタをかけてもよい。

【００５１】次に、入力信号生成部４－２ａでは、生成した音源信号が空中を伝播してマイクロホン位置に到達すると仮定したときのマイクロホンで観測される信号を計算する。マイクロホン位置で観測される信号は、音源信号の伝搬時間差と伝搬に伴う振幅変化から計算できる。

【００５２】例えば、マイクロホンと人物位置が図７のような設定であるとして、図７を参照して次のように行う。図７（ａ）のように、１番目のマイクロホンの座標を（x_1，y_1）、ｉ番目のマイクロホンの座標を（x_i，y_i）とすると、平面波を仮定した場合、θ方向から音波が入射する際のｉ番目のマイクロホンと１番目のマイクロホンに入射する音波の伝搬時間差τ_i は、

☆

$$\tau_i(\theta) = ((x_i - x_1)^2 + (y_i - y_1)^2)^{1/2} \times \cos(\theta - \tan^{-1}((y_i - y_1)/(x_i - x_1))) \quad (6)$$

$$\text{振幅は} a_1 = a_2 = \cdots a_N = 1 \quad (7)$$

とおくことができ、点音源を仮定した場合、図７（ｂ）◇　◇のように仮想音源位置θを（x_s，y_s）とおくと

$$\tau_i = ((x_i - x_s)^2 + (y_i - y_s)^2)^{1/2} - ((x_1 - x_s)^2 + (y_1 - y_s)^2)^{1/2} / c \quad (8)$$

$$\text{振幅は} a_i = ((x_i - x_s)^2 + (y_i - y_s)^2)^{1/2} / ((x_1 - x_s)^2 + (y_1 - y_s)^2)^{1/2} \quad (9)$$

となる。ただし、ｃは音速である。なお、ここでは簡略化のため２次元平面上で説明したが、３次元空間への拡張は容易である。

【００５３】上のようにして求めた遅延時間τ_i を用　＊い、ｉ番目のマイクロホンの観測信号のうち、ｋ番目の人物位置から到来する成分$x_i(n)$は、$S_k(n)$をｋ番目の人物位置から到来する音波の波形、または音源信号として、

$$x_i(n) = S_k(n - \tau_i') \quad (10)$$

15

とできる。ここで、τ_i ′はτ_i を四捨五入した値である。また、信号の遅延をもっと精度よく行うため、四捨五入する代わりに、上記した音響システムとデジタル処理（ｐｐ．２１５）に述べられているようにデジタルフィルタを畳み込んでもよいし、フーリエ変換により周波数領域に変換して位相回転により遅延を与えた後、逆フーリエ変換してもよい。

【００５４】次に、学習信号レベル計算部４－２ｃにおいて音源信号のレベルを決め、以上のようにして求めたマイクロホン位置での音源信号の観測値が、決定したレベルと一致するようにして、実際にマイクロホンから入力する信号とこの音源信号を加算し、適応フィルタの入力信号として出力するようにする。音源信号レベルの計＊

$$A_k = (P_N * 10^{v/10})^{1/2}$$

により計算する。ここで、A_k は音源信号の振幅、P_N は背景雑音のパワーである。

【００５６】次に、希望応答生成部４－２ｄでは、前記の２つの抑圧処理モード（Ａ），（Ｂ）に応じて、別の方法で希望応答を生成する。選択された人物位置すべてから到来する音波を収集する場合（Ａ）は、選択されたすべての人物位置から音波が到来すると仮定したときのマイクロホン位置での観測信号を希望応答として出力するようにする。この場合、例えば、１番目のマイクロホン位置で観測される信号を希望応答として使うようにする。ただし、マイクロホン位置で観測される信号よりも遅延させたものとするようにする。遅延の大きさは、例えばタップ数の半分とする。

【００５７】選択された人物位置のうち、ある人物位置からの音波について抑圧したい場合（Ｂ）では、人物位置から音波が到来すると仮定したときのマイクロホン位置での観測信号作成の際に、その人物位置からの音波に相当する信号は加算しないようにする。例えば、選択した人物位置が３個で、入力したい人物位置がその中の１個の場合は、入力したい１個の人物位置からの到来だけを仮定してマイクロホンで観測される信号を希望応答とする。

※

$$P_N = \gamma * P_N + (1-\gamma)\, P_N' \qquad (12)$$

ここで、P_N ′は、それまでに求まっていた背景雑音パワー、γは忘却係数であり、例えば、$\gamma = 0.1$である。

【００６１】次に、学習信号レベル計算部４－２ｃにおいて、音源信号が伝播してマイクロホン位置で観測されると仮定したときの信号を計算し、式（１１）により音★

$$x_i(n) = r_i(n) + \Sigma A_k\, S_k\, (n - \tau_k') \qquad (13)$$

により計算する（ステップＳ４）。次に、希望応答生成部４－２ｄにおいて、音源信号と音源信号の振幅から☆

$$d(n) = \Sigma A_k\, S_k\, (n - \tau_k' - n_o) \qquad (14)$$

となる。ここで、n_o は適当な遅延、例えば、$n_o = 10$である。ただし、抑圧処理のモードがＡの場合、ｋはすべての人物位置について変化させ、モードがＢの場

16

＊算のため、学習信号レベル計算部４－２ｃにおいては、入力音声の背景雑音パワーの音声区間の平均値を計算して保持するようにする。入力音声の背景雑音パワーは、例えば複数ある中の１番目のマイクロホンのパワーを逐次計算して音声区間を検出し、音声区間として検出されなかった区間の平均パワーを求めるようにする。パワーに基づいた音声区間検出はよく知られているように、例えば文献（新美著：音声認識）に詳述されている。

【００５５】このようにして求めた背景雑音パワーに対して一定値ｖ、例えば、ｖ＝７ｄＢ高い値を音源信号レベルとし、このパワーの平方根の値を音源信号の振幅とするようにする。すなわち、

（１１）

※【００５８】以上に述べた環境適応部４－２を含む雑音抑圧部４における音声など連続信号の処理は例えば、１ｃｈあたり１０２４点を１ブロックとし、ブロック単位で行うようにする。すなわち、音声入力部３における音声データの読み込み、環境適応部４－２における音源信号と学習信号の生成、適応フィルタ処理、雑音除去処理などは、すべて１ｃｈあたり１０２４点を１ブロックとしてブロック単位で行うものとする。

【００５９】ここで、以上に述べた環境適応部４－２の処理の流れについて図８を参照しながら説明する。

【００６０】まず、環境適応部４－２の音源信号生成部４－２ｂにおいて、選択人物位置の数の系列の音源信号を生成する（ステップＳ１）。音源信号は音源ごとに無相関な系列とし、分散は１に正規化しておくようにする。次に、学習信号レベル計算部４－２ｃにおいて、複数チャネルで入力した入力音声の中から、例えば１ｃｈ目の信号のパワーを、例えば波形１２８点の小セグメントごとに計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーを求め、これを背景雑音パワーP_N とする（ステップＳ２）。背景雑音パワーは、それまでに求まっていた値との間で平均化してもよく、その場合、次式により平均化する

★源信号の振幅A_k を計算する（ステップＳ３）。次に、入力信号生成部４－２ａにおいて、実際の入力音声と加算して適応フィルタの入力信号を生成する。すなわち、ｉｃｈ目の入力音声を$r_i(n)$とすると、ｉｃｈ目の適応フィルタの入力信号$x_i(n)$は、

☆希望応答を生成する。式で表すと、

合、ｋは感度を高く設定する人物位置について変化させる。式（１２）（１３）のように、音源信号の遅延をサンプリング周期で四捨五入した値τ_k ′により遅延を与

えるかわりに、もっと精度よく遅延させることも可能であることはすでに述べた（ステップＳ５）。

【００６２】次に、生成した入力信号と希望応答を適応フィルタに入力し、フィルタ係数を得る（ステップＳ６）。得られたフィルタは、雑音除去部４－３に送り、入力音声を処理して音声を抽出する。雑音除去部４－３におけるフィルタ処理は、式（４）に従って行う。

【００６３】次に、図９を参照しながら本実施形態全体の処理の流れを説明する。

【００６４】まず、初期設定を行い、選択する人物位置の数Ｎと雑音抑圧処理のモードＡかＢかを設定する（ステップＳ３１）。

【００６５】画像の処理の方では、画像データをビデオカメラ５から、例えば毎秒５フレームで取り込み（ステップＳ３２）、フレームごとに人物位置を特定して出力し（ステップＳ３３）、このステップＳ３２とＳ３３を繰り返す。画像から人物位置を特定する画像の処理は、音声の雑音抑圧処理とは独立に、並列に処理するようにする。

【００６６】音声処理の方では、まず、音声データを、例えばサンプリング１２ｋＨｚでＡＤ変換し、１チャネルあたり、例えば１０２４サンプルを１ブロックとして１ブロック分のデータを取り込む（ステップＳ３４）。次に、人物位置が特定されているか否かを判定し（ステップＳ３５）、人物位置が特定されていない場合は、何もせずにこのステップＳ３４とＳ３５を繰り返し、特定された場合は次のステップＳ３６に進む。人物位置は、処理開始直後で画像処理結果が出ていない場合や人物がいない場合に特定されない。位置画像に関する処理と音声に関する処理とは独立しているため、例えば、一つの計算機上で全処理を行う場合、よく知られているように、ソケットを用いたプロセス間通信やシェアドメモリ、あるいはファイルを通じて人物位置のデータのやり取りを行うことができる。

【００６７】次に、ステップＳ３６では、人物位置選択部４－１において、処理対象とする人物位置を選択する。次に、環境適応部４－２において、人物位置選択部４－１で選択された人物位置または方向と距離を用いて適応フィルタの学習信号を生成し、フィルタ係数を更新する（ステップＳ３７）。学習信号の長さは、取り込んだ音声データの長さと同じく１ｃｈあたり１０２４点にする。

【００６８】次に、ステップＳ３７で更新されたフィルタ係数を雑音除去部４－３にコピーし、このフィルタと入力音声との畳み込み演算を行って音声を出力する（ステップＳ３８）。以上のステップＳ３１からＳ３２の処理とＳ３３からＳ３８までの処理を並列に繰り返す。

【００６９】以上に述べた処理により、画像処理により特定された複数の人物位置各々から到来する音声の感度を設定できるように雑音抑圧処理を行うフィルタの係数を学習しているため、複数の人物が同時に発声した場合に、その人物すべての音声を背景雑音を抑圧して取り出したり、一人だけの人物の音声のみを他の人物の音声を抑圧して取り出すことが可能となる。

【００７０】また、逐次人物位置を特定し、その人物位置に応じてフィルタ処理の学習信号を生成しているため、複数の人物が各々動いた場合でも追随して雑音抑圧処理を行うことが可能である。

【００７１】以下に本発明の第２実施形態を説明する。第２実施形態では、音声処理の対象として選択された人物各々からの発声音を検出し、この検出情報に基づいて学習信号の生成を制御することによって、学習を高精度に行う。

【００７２】第１実施形態で述べたパイロット信号法による適応フィルタの学習では、人物が発声中であるか否かに関わらず学習を行っていたが、感度を高くして入力したい人物が発声している間は適応を止めることにより、また、抑圧したい人物の発声中は、その人物方向からの到来を仮定した音源信号を使わずに入力信号と希望信号を生成することにより、より環境に適応した高精度な雑音抑圧の学習が行える。このため、本実施形態では、人物位置ごとに発声中であるかどうかの目安となる位置別の到来パワーを推定する位置別到来パワー推定部４－４をさらに具備しており、これを含んだ全体構成を図１０に示す。図１０において、４－１は人物位置選択部、４－２は環境適応部、４－３は雑音除去部である。

【００７３】また、推定した位置別到来パワーに基づいてフィルタ学習の制御を行うため、環境適応部４－２は図１１のような構成を具備している。図１１において、４－２ａは適応フィルタの入力信号生成部、４－２ｄは適応フィルタの希望応答生成部、４－２ｅは適応フィルタによる適応処理部、４－２ｂは入力信号と希望応答生成の際に用いる人工的な波形である音源信号を発生する音源信号生成部、４－２ｃは入力信号と希望応答の生成の際、人物位置ごとの音源信号の振幅を位置別到来パワーに基づいて決定する学習信号レベル計算部、４－２ｆは位置別到来パワーからフィルタ学習の際の適応速度を制御するパラメータを決定する適応制御信号制御部である。

【００７４】位置別到来パワー推定部４－４では、マイクロホンアレイ６に入力した音声から、人物位置ごとの到来パワーを求める。マイクロホンアレイ６によって位置あるいは方向ごとの到来パワーを計算する方法としては、文献（音響システムとデジタル処理）に詳述されているように、遅延和法、最小分散法、ＭＵＳＩＣ法など種々の方法があるが、ここでは、少ない計算量で実現可能な遅延和法による方法を説明する他の方法も計算量が多くなるだけで適用可能であることは言うまでもない。

【００７５】上述の文献にも詳述されているように、遅延和法は、各マイクロホンからの信号を対象とする方向

または位置から到来する音波の位相が揃うように遅延させてから和をとるものである。図２に示すようなマイクロホンと到来位置の関係の場合、ｉ番目のマイクロホンと１番目のマイクロホンに入射する音波の伝搬時間差τｉは、平面波が入射する場合は式（６）により、球面波が入射する場合は式（８）により計算できる。このと ＊

$$p=|\Sigma x_i\ (n-\tau_i)\ |^2 /M \qquad (15)$$

であり、この値は、対象とする方向または位置から音波が到来している場合には音源のパワーに比例することが知られているので、式（１５）により、各人物位置から到来するパワーが推定できる。なお、球面波の場合は音源とマイクロホンとの距離により補正係数が必要になるが、容易に補正できる。詳細は多数センサによる音源波形推定に関する文献（日本音響学会誌、４７、４、ｐｐ２６８－２７３、１９９１）に述べられている。

【００７６】次に、学習信号レベル計算部４－２ｃについて説明する。ここでは、求められた位置別到来パワーから、適応フィルタの入力信号と希望応答の生成の際の、各々の人物位置の音源信号の振幅を決定する。このため、学習信号レベル計算部４－２ｃでは、入力音声の背景雑音パワーと位置別到来パワーの音声区間の平均値を計算して保持するようにする。入力音声の背景雑音レベルは、複数のマイクロホン中の、例えば１番目のマイクロホンのパワーを逐次計算して音声区間を検出し、音声区間として検出されなかった区間の平均パワーを求め※

$$A_k=(P_N * 10^{v/10})^{1/2} \qquad (16)$$

（位置別到来パワーが背景雑音＋ｖｄＢより小さいと ★ ★き）

$$A_k=A_{ko}=(P_k)^{1/2} \qquad (17)$$

（位置別到来パワーが背景雑音＋７ｄＢより大きいとき）のように計算する。ここで、P_N は背景雑音のパワー、P_k はｋ番目の位置の位置別到来パワーである。 ☆

$$A_k=A_{ko}*(P_N/P_k)^{1/2} \qquad (18)$$

によって計算する。以上のようにして求めた音源振幅と入力音声とを加算し、適応フィルタの入力信号を生成する。

【００８０】例えば、選択された人物位置がａとｂの２個であり、ａが音声入力の対象とする目的の人物位置であり感度を高く設定する位置、ｂが感度を低く設定する妨害音の位置であるとする。図１２の（１）、（２）に示すようにａ，ｂの位置ごとの到来パワーが推定された場合、１，２で示した区間では、（３）に示すように入力信号中のａに関する成分を大きくし、３で示した区間では（４）に示すようにｂに関する成分を小さくする。また、希望応答は、入力信号中のａに関する成分と同じとし、ｂに関する成分はすべて０とするか加算しないようにする。 ◆

$$a=C/(\alpha_B\ p_k/p_N+1) \qquad (19)$$

（抑圧モードＢのとき）

$$a=C/(\alpha_A\ \Sigma(p_k/p_N)/M+1) \qquad (20)$$

（抑圧モードＡのとき）

ここで、P_N は背景雑音のパワー、P_k はｋ番目の位置の位置別到来パワー、$C\alpha_A$、α_B は定数、例えば、Ｃ

＊き、ｉ番目のマイクロホンの波形を$x_i(n)$とし、時刻ｎにおいてＪサンプル過去から時刻ｎまでの各マイクロホンの波形サンプルの系列$x_i=(x_{i(n-J+1)}, x_{i(n-J+2)}, \cdots, x_{i(n-1)}, x_{i(n)})$をτｉ遅延させた場合の全マイクロホンについての平均パワーは、

※るようにする。パワーに基づいた音声区間検出はよく知られているように、例えば文献（新美著：音声認識）に詳述されている。また、位置別到来パワーに関しても同様に位置ごとに音声区間の検出を行い、こちらは音声区間の平均パワーを求めるようにする。

【００７７】このようにして求めた位置別到来パワーと背景雑音パワーから、音源信号の振幅を計算する。このとき、人物位置が、感度を高く設定した位置であるか、低く設定した位置であるかに応じて振幅の計算方法を変えるようにする。

【００７８】感度を高くするように設定した位置の場合は、上記のように求めた位置別到来パワーの位置ごとの平均値の平方根の値を音源信号の振幅とするようにする。なお、発声がない場合は、位置別到来パワーは小さい値となるため、位置別到来パワーが背景雑音に対してある値ｖ、例えば、ｖ＝７ｄＢを上回る時だけ位置別到来パワーの平方根の値に設定するようにする。すなわち、A_k をｋ番目の位置の音源信号の振幅とすると、

☆【００７９】感度を低くするように設定した位置の場合、位置別到来パワーが大きいほど小さい振幅となるようにする。例えば、ｋ番目の位置の音源信号の振幅を、

◆【００８１】次に、位置別到来パワーに基づいた適応フィルタの適応速度の制御について説明する。適応フィルタの学習は、よく知られているように、式（５）のステップＳサイズの値ａにより制御できる。ここでは、音声入力の対象となる位置からの到来パワーが大きい場合は、入力信号の中の抽出すべき信号があるにも関わらず希望応答の中にその信号がないため、抑圧の対象となってしまう。そこで、この到来パワーの値が大きいときは適応を遅くあるいは停止し、小さいときは適応を早くするように適応速度を制御する。

【００８２】このため、例えば、式（５）で固定していたステップサイズ（式（５）のａ）の値を次式により逐次計算して可変とするようにする。

【００８３】

＝２．０、$\alpha_A = \alpha_B = 1$である。ステップサイズの計算式として挙げた上式は一例であり、他の方法も使用可能である。

【００８４】ここで、図１３を参照しながら第２実施形態の環境適応部全体の処理の流れを説明する。

【００８５】まず、環境適応部４－２の音源信号生成部４－２ｂにおいて、選択人物位置の数の系列の音源信号を生成する（ステップＳ１１）。 次に、学習信号レベル計算部４－２ｃにおいて、複数チャネルで入力した入力音声のパワーを計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーから背景雑音パワーP_Nを求める（ステップＳ１２）。このとき、式（１２）により平均化してもよい。次に、複数チャネルの入力音声から式（１５）により位置別到来パワーを計算する（ステップＳ１３）。

【００８６】次に、学習信号レベル計算部４－２ｃにおいて、式（１６）から式（１８）により音源信号の振幅A_kを計算する（ステップＳ１４）。次に、入力信号生成部４－２ａにおいて、式（１３）により、実際の入力音声と加算して適応フィルタの入力信号を生成する（ステップＳ１５）。 次に、希望応答生成部４－２ｄにおいて、式（１４）により、音源信号と音源信号の振幅から希望応答を生成する（ステップＳ１６）。次に、適応制御信号生成部４－２ｆにおいて、式（１９）または（２０）により、背景雑音パワーと位置別到来パワーからステップサイズの系列である適応制御信号を生成する（ステップＳ１７）。

【００８７】次に、生成した入力信号と希望応答と適応制御信号を適応フィルタに入力し、フィルタ係数を得る（ステップＳ１８）。得られたフィルタは、雑音除去部４－３に送り、入力音声とフィルタを畳み込んで音声を抽出する。雑音除去部４－３におけるフィルタ処理は、式（４）に従って行う。

【００８８】第２実施形態の全体の処理の流れは第１実施形態と同じであるので改めて述べない。

【００８９】以上に述べたように、画像処理により特定された複数の人物位置各々から到来する音声に対し、適応フィルタにより感度を設定して雑音抑圧を行う際、人物位置からの到来音のパワーに応じて適応フィルタの適応処理を制御しているため、実環境に応じた高精度な適応が行え、雑音抑圧性能を大幅に高くしながら、複数の人物が同時に発声した場合に、その人物すべての音声を背景雑音を抑圧して取り出したり、一人だけの人物の音声のみを他の人物の音声を抑圧して取り出すことが可能となる。

【００９０】また、画像から逐次人物位置を特定し、その人物位置に応じてフィルタ処理の学習信号を生成して＊いるため、複数の人物が各々動く場合でも追随して雑音抑圧処理を行うことができる。

【００９１】以下に、音声パワー検出と画像からの発声動作検出を行なう第３実施形態について説明する。第３実施形態は、第２実施形態において行っていた人物位置ごとの到来パワー推定に加え、画像データに基づいた発声動作の検出を行い、これら２つの情報に基づいて学習信号の生成と適応速度の制御を行うことにより、音の環境をより正確に反映して適応フィルタの学習を行うようにする。本実施形態では画像から発声動作を検出しているため、高雑音下でも人物が発声中かどうかを精度よく検出でき、高精度な適応フィルタの学習の制御が行える。

【００９２】画像に基づいた発声動作の検出と位置別到来パワーに基づいて適応フィルタの学習を制御するため、本実施形態の雑音抑圧部は、第２実施形態の雑音抑圧部の構成にさらに画像から発声動作に関する情報を検出する発声動作情報検出部を追加し、図１４のような構成としている。

【００９３】図１４において、１はビデオカメラなどから画像を入力する画像入力部、２は入力した画像を処理して人物の位置を特定する人物位置検出部、３は複数のマイクロホンからの音声を並列に入力する音声入力部、４は画像処理により検出された人物位置に基づいて複数のマイクロホンからの入力音声から雑音を抑圧して音声を取り出す雑音抑圧部である。

【００９４】雑音抑圧部４は、人物位置検出部２により特定された人物位置の中から処理対象とする人物位置を選択する人物位置選択部４－１と、選択された人物位置に基づいて適応フィルタの学習を行う環境適応部４－２と、決定されたフィルタ係数により雑音抑圧処理を行う雑音除去部４－３と、人物位置ごとに到来パワーを検出する位置別到来パワー推定部４－４と、人物位置ごとに画像から発声動作に関する情報を検出する位置別発声動作情報検出部４－５とからなる。

【００９５】画像による音声区間の検出は、口元の画像の時間変化から行う方法が知られており、簡単には、口元画像全体の輝度変化を時刻ごとに計算し、その変化が大きい時刻を発声中であるとして検出できる。ここでは、正確な口元画像の代わりに人物位置検出部２で特定した人物の顔を含む画像において、例えばその下半分についての画面全体にわたる輝度の時間変化を求めて発声動作の目安とするようにする。画像データのフレームの番号をｉ、縦横位置ｘ，ｙにおける人物位置の顔を含む顔周辺画像データをＧ（ｉ，ｘ，ｙ）とすると、フレームｉとｉ－１の間の輝度変化は、

$$K(i) = \sum_x \sum_y |G(i, x, y) - G(i-1, x, y)|^2 \qquad (20-1)$$

と計算でき、このＫ(ｉ) の値を発声動作があるか否かの目安とする。人物の顔周辺画像は、人物位置検出部２か

ら、特定した位置とともに入力するようにする。人物位置特定の処理では、全体画像中から顔の部分の画像を切り出す処理を含むのが一般的であるため、顔周辺画像は容易に取り出せる。画像からの発声動作の検出方法は輝度変化の計算に限るものではなく、他の方法も使用可能である。以降上の輝度変化Ｋ(i) を含め、画像から抽出した発声動作の目安となる情報を便宜上発声動作情報と呼ぶことにする。

【００９６】画像からの人物位置検出処理の速さは、画像の入力レート、例えば、５フレーム／秒で行うので、音声処理をブロック単位で行う場合の処理速度とは一致せず音声処理より遅いのが普通である。従って、位置別発声動作情報検出部４－５に入力する画像は、音声処理に関する１ブロック前と同じものを使う場合があることになるが、その場合、同じ画像間の輝度変化を求めることになるので輝度変化は０になる。この状況を避けるため、輝度変化の計算の結果、値が０のときは１ブロック前の輝度変化の値をそのまま出力するようにする。

【００９７】発声動作情報は、位置別到来パワーと並列に使うようにしており、環境適応部４－２の学習信号レベル計算部４－２ｃと、適応制御信号生成部４－２ｆにおいて用いている。本実施形態の他の部分は第２実施形＊態と同じであり、環境適応部４－２の構成も同じであるので、学習信号レベル計算部４－２ｃと適応制御信号生成部６についてのみ述べる。

【００９８】まず、学習信号レベル計算部４－２ｃでは、適応フィルタの入力信号と希望応答の生成の際の、到来を仮定する人物位置各々の音源信号の振幅を決定する。このため、第２実施形態と同様、学習信号レベル計算部４－２ｃでは、入力音声の背景雑音パワーと位置別到来パワーの音声区間の平均値を計算して保持し、上述の発声動作情報と、求めた位置別到来パワーと背景雑音パワーとから、音源信号の振幅を計算する。このとき、音源の存在を仮定する人物位置が、感度を高く設定した位置であるか、低く設定した位置であるかに応じて振幅の計算方法を変えるようにする。

【００９９】感度を高くするように設定した位置の場合、第２実施形態と同様、位置別到来パワーの平均値の平方根の値を音源信号の振幅とするようにする。なお、発声がない場合は、位置別到来パワーは小さい値となってしまうため、背景雑音に対してある値ｖ、例えばｖ＝５ｄＢ高い値を最小値として設定し、位置別到来パワーがこれを上回る時だけ検出した値に設定するようにする。すなわち、

$$A_k = (P_N * 10^{v/10})^{1/2} \quad (21)$$

（位置別到来パワーが背景雑音＋ｖｄＢより小さいと※※き）

$$A_k = (P_k)^{1/2} \quad (22)$$

（位置別到来パワーが背景雑音＋ｖｄＢより大きいとき）のように計算する。ここで、P_N は背景雑音パワーの平均値、P_k は位置別到来パワー、ｖは最小値５である。

【０１００】感度を低くするように設定した位置の場★合、位置別到来パワーと発声動作情報が大きいほど音源信号が小さい振幅となるようにして、人工的な学習信号への適応を弱めるようにする。例えば、ｋ番目の音源信号の振幅を、

$$A_k = A_k * \gamma (P_N / P_k)^{1/2} * (1-\gamma) (K_O / (K_k + K_O))^{1/2} \quad (23)$$

によって計算する。ここで、K_k は式（２０）によって計算したｋ番目の位置の顔周辺画像のフレーム間の輝度変化、K_O は同輝度変化の平均値、γは定数、例えば、γ＝０．５とする。以上のようにして求めた音源振幅と入力音声とを加算し、適応フィルタの入力信号を生成する。

【０１０１】次に、適応制御信号生成部４－２ｆは、位置別到来パワーと発声動作情報に基づいてフィルタ学習のステップサイズの制御を行う。ここでは、実施例２と☆同様、この到来パワーの値が大きいときは適応を遅く、小さいときは適応を早くするようにステップサイズを制御する。

【０１０２】このため、例えば、式（５）で固定していたステップサイズ（式（５）のａ）の値を次式により逐次計算して可変とすることにより、適応の速度を調整するようにする。

【０１０３】

$$a = C / (\alpha_B p_k / p_N + \beta_B K_k + 1) \quad \text{（抑圧モードＢ）} \quad (24)$$

$$a = C / (\alpha_A \Sigma (p_k / p_N) + \beta_A \Sigma (K_k / K_O) + 1) \quad \text{（抑圧モードＡ）} \quad (25)$$

ここで、p_k は感度を高くするように設定した位置ｋからの位置別到来音パワー、Ｃ、α_A、α_B、β_A、β_B は定数、例えばＣ＝２．０、$\alpha_A = \alpha_B = 0.5$、$\beta_A = \beta_B = 0.5$ である。ステップサイズの計算式として挙げた上式は一例であり、例えば、位置別到来パワーと輝度変化の値に対して各々しきい値を定め、どちらか一方がこれを越えた場合に、適応を止める（ａ＝０とする）など、他の方法も使用可能である。

【０１０４】ここで、図１５を参照しながら第３実施形態の環境適応部全体の処理の流れを説明する。

【０１０５】まず、環境適応部４－２の音源信号生成部４－２ｂにおいて、選択人物位置の数の系列の音源信号

を生成する（ステップＳ２１）。

【０１０６】次に、学習信号レベル計算部４－２ｃにおいて、複数チャネルで入力した入力音声のパワーを計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーから背景雑音パワーP_Nを求める。このとき、式（１２）により平均化してもよい（ステップＳ２２）。

【０１０７】次に、複数チャネルの入力音声から式（１５）により位置別到来パワーを計算する（ステップＳ２３）。次に、位置別発声動作情報検出部４－５において、人物位置ごとの顔周辺画像を人物位置検出部２から入力し、発声動作情報を検出する。輝度変化の計算の結果、値が０ならば１ブロック前の値をこのブロックの輝度変化の値とし、０以外なら計算結果をこのブロックの輝度変化の値とし、この値を記憶する（ステップＳ２４）。

【０１０８】次に、学習信号レベル計算部４－２ｃにおいて、式（２１）から（２３）により音源信号の振幅A_kを計算する（ステップＳ２５）。次に、入力信号生成部４－２ａにおいて、式（１３）により、実際の入力音声と加算して適応フィルタの入力信号を生成する（ステップＳ２６）。次に、希望応答生成部４－２ｄにおいて、式（１４）により、音源信号と音源信号の振幅から希望応答を生成する（ステップＳ２７）。

【０１０９】次に、適応制御信号生成部４－２ｆにおいて、式（２４）または（２５）により、背景雑音パワーと位置別到来パワーと位置別発声動作情報とからステップサイズの系列である適応制御信号を生成する（ステップＳ２８）。次に、生成した入力信号と希望応答と適応制御信号を適応フィルタに入力し、フィルタ係数を得る（ステップＳ２９）。得られたフィルタは、雑音除去部４－３に送り、入力音声とフィルタを畳み込んで音声を抽出する。

【０１１０】第３実施形態の全体の処理の流れは第１実施形態と同じであるので改めて述べない。以上に述べたように、画像処理により特定された複数の人物位置各々から到来する音声に対し、適応フィルタにより感度を設定して雑音抑圧を行う際、人物位置からの到来音のパワーと画像から求めた発声動作情報に応じて適応フィルタの適応処理を制御しているため、雑音が大きく、位置別到来パワーの推定が低い場合でも、雑音抑圧性能を大幅に高くしながら、複数の人物が同時に発声した場合に、その人物すべての音声を背景雑音を抑圧して取り出したり、一人だけの人物の音声のみを他の人物の音声を抑圧して取り出したりできる。

【０１１１】以下にテスト発声モードを備えた第４実施形態について説明する。第４実施形態は、音声収集装置の動作中に、一時、通常の音声入力処理を停止し、テスト発声を入力してレベル計算を行うための、入力モード切り替え部を具備することにより、音源信号のレベルを実環境の値に合わせ、高精度の適応処理を行うものである。

【０１１２】これまでに述べた実施形態では、適応フィルタの学習信号の生成に用いる音源信号は、背景雑音レベルと経験的に決めたデフォルトの音声のレベル値を用いてその振幅を計算してきたが、本実施形態では、さらに現実の音場に正確に合わせるため、テスト発声を行って音源のレベルを決めるようにしている。このため、テスト発声か、通常の音声入力かを動作中に切り替える入力モード切り替え部７を追加し、図１６に示すような構成としている。図において、１は画像入力部、２は人物位置検出部、３は音声入力部、４は雑音抑圧部、５は入力モード切り替え部である。

【０１１３】この構成において、通常は入力モード切替部には、通常の音声入力処理であることを設定しておき、テスト発声時には入力モード切替部から、テスト発声であることを入力する。入力モードをテスト発声に設定した場合、通常行っている適応フィルタ処理は止め、学習信号レベル計算部４－２ｃにおいて、入力音声のレベルを計算し、保持するようにする。テスト発声を終了して通常の入力モードに戻った際は、学習信号レベル計算の際、デフォルトで例えば５ｄＢなどと決めてきた音源信号の最小値を使わず、ここで求めたテスト発声のレベルから音源信号の振幅を計算するようにする。

【０１１４】ここで、図１７を参照しながら第４実施形態の全体の処理の流れを説明する。

【０１１５】まず、初期設定を行い、選択する人物位置の数Ｎと雑音抑圧処理のモードＡかＢかを設定する（ステップＳ４１）。

【０１１６】画像の処理の方では、画像データをビデオカメラから、例えば毎秒５フレームで取り込み（ステップＳ４２）、フレームごとに人物位置を特定して出力し（ステップＳ４３）、このステップＳ４２とＳ４３を繰り返す。画像から人物位置を特定する画像の処理は、音声の雑音抑圧処理とは独立に、並列に処理するようにする。

【０１１７】音声処理の方では、まず、音声データを、例えばサンプリング１２ｋＨｚでＡＤ変換し、１チャネルあたり、例えば１０２４サンプルを１ブロックとして１ブロック分のデータを取り込む（ステップＳ４４）。

【０１１８】次に、ステップＳ４５で入力モードがテスト発声か通常入力かを検査し、テスト発声であればステップＳ４６に進み、通常入力であればステップＳ４７に進む。ステップＳ４６では、学習信号レベル計算部４－２ｃにおいて、入力音声のレベルを計算して保持する。入力音声のレベルは、ある番号、例えば１ｃｈ目のマイクロホンからの入力のパワーに基づいて音声検出を行い、音声区間として検出された部分の平均値を用いるようにする。この後、ステップＳ４４に戻る。

【０１１９】次に、ステップＳ４７では人物位置が特定

されているか否かを判定し、人物位置が特定されていない場合は、何もせずにこのステップＳ４４乃至Ｓ４７を繰り返し、特定された場合は次のステップＳ４８に進む。人物位置は、処理開始直後で画像処理結果が出ていない場合や人物がいない場合に特定されない。位置画像に関する処理と音声に関する処理とは独立しているため、例えば、一つの計算機上で全処理を行う場合、よく知られているように、ソケットを用いたプロセス間通信やシェアドメモリ、あるいはファイルを通じて人物位置のデータのやり取りを行うことができる。

【０１２０】次のステップＳ４８では、人物位置選択部４－１において、処理対象とする人物位置を選択する。次に、環境適応部４－２において、人物位置選択部４－１で選択された人物位置または方向と距離を用いて適応フィルタの学習信号を生成し、フィルタ係数を更新する（ステップＳ４９）。次に、ステップＳ４９で更新されたフィルタ係数を雑音除去部４－３にコピーし、このフィルタと入力音声との畳み込み演算を行って音声を出力する（ステップＳ５０）。

【０１２１】以上のステップＳ４１からＳ４２の処理とステップＳＳ４３からＳ５０までの処理を並列に繰り返す。

【０１２２】なお、本実施例で述べたテスト発声モードは、第１実施形態に追加する形で述べたが、第２、第３実施形態に述べた構成に追加して併用することも可能である。

【０１２３】次に、本発明の第５実施形態について説明する。本実施形態では、人物ごとの音声を他の人物の音声と分離して取り出すため、第１乃至第４実施形態の雑音除去部と、環境適応部における適応処理部各々を複数のフィルタから構成するようにしている。これを、図１８に示す。この部分以外については第１乃至第４実施形態と同様の構成である。なお、この実施形態は、第３実施形態の拡張として説明するが、第２、第４実施形態にも適用可能であり、また、環境適応部の適応制御信号生成部を取り去るだけで第１実施形態を拡張した場合にもなっている。図１８において、環境適応部４－２における適応処理部４－２ｅと雑音除去部４－３におけるフィルタは各々複数個（Ｎ個）づつ、例えば、３個づつ用意し、雑音除去部４－３のフィルタは環境適応部４－２の適応フィルタの係数のコピーである。また、環境適応部４－２の適応フィルタには、すべて同じ入力信号を入力するが、希望応答と適応制御信号は、フィルタの番号ｋにより異なった内容のものを入力するようにする。

【０１２４】この適応処理部４－２ｅを含む環境適応部４－２の処理について、次に説明する。まず、環境適応部４－２では、画像により検出され選択された人物位置各々から音波が到来すると仮定し、その音波の信号を音源信号生成部４－２ｂで生成する。この信号の内容は、人工的なもの、例えば音源間で無相関なランダム雑音で良いことは、第１実施形態で述べた。この音源信号をもとに、適応処理を行うための入力信号と希望応答を生成する。その際、学習信号レベル計算部４－２ｃにおいて、位置別到来パワー、発声動作情報のいずれか又は両方と観測した背景雑音レベルとに基づいて音源信号の振幅を決定する。

【０１２５】また、適応処理の際の適応速度の制御を行う適応制御信号を適応制御信号生成部４－２ｆにおいて生成する。適応処理部４－２ｅでは、上記３つの信号を入力として適応フィルタにより雑音抑圧のためのフィルタ係数を決定する。なお、適応制御信号は必ずしも必要でなく、また、学習信号レベルの計算には、位置別到来パワーと発声動作情報は必ずしも必要でない。

【０１２６】人物位置検出部２により複数検出され、その中から人物位置選択部４－１により選択された複数の人物位置各々から到来する音声の抽出を、構成図で示したように、複数個のフィルタを使って行う。フィルタの数は選択人物位置の数と一致するようにし、フィルタの番号ｋは人物位置の番号に対応させるようにする。

【０１２７】各フィルタが各人物の音声を抽出するようにするため、ｋ番目の適応フィルタによる適応の際の希望応答の内容は、ｋ番目の人物位置から到来することを仮定する１個の音源の信号と同じになるようにし、式（１４）により計算される。また、各適応フィルタの入力信号は、Ｎ個の人物位置から各々に対応する音源信号が伝搬してマイクロホン位置で観測されるときの信号をすべて重ね合わせたものに実際に入力した音声を加算したものであり、式（１３）により、マイクロホン位置ごとに計算されてＮチャネルの信号が生成される。入力信号はすべての適応フィルタで共通に使われる。

【０１２８】一方、適応フィルタの収束速度を制御する適応制御信号は、適応フィルタの番号ごとに異なった信号内容のものを生成するようにし、ｋ番目のフィルタには、ｋ番目の人物位置からの位置別到来パワーまたは発声動作情報に基づいて式（１９）、（２０）または式（２４）、（２５）により計算されたステップサイズの値の系列を入力するようにする。位置別到来パワーまたは発声動作情報が得られない第１実施形態を拡張する場合は、適応制御信号は生成せず、ステップサイズは一定値とする。

【０１２９】上記のようにして生成した入力信号、希望応答、適応制御信号を適応フィルタに入力して複数組のフィルタ係数を決定した後、これらの係数を雑音除去部４－３に送り、入力音声をフィルタ処理して雑音を除去し、各人物の音声を別々に抽出するｋ番目の人物位置の音声はｋ番目のフィルタから出力されることになる。

【０１３０】以上のように、人物位置に対応した複数のフィルタを用いることにより、各人物位置からの到来音を別々に分離して取り出すことが可能となる。

【０１３１】次に、本発明の第６実施形態について説明

する。第６実施形態はパイロット信号法による適応フィルタでなく、線形拘束条件付き適応フィルタにより、雑音抑圧処理を行って音声収集するものである。この種類の適応フィルタにより計算量の多い学習信号の生成処理を省いた処理が可能である。

【０１３２】図１９は第６実施形態の全体構成を示す図である。図１９において、１は画像を入力する画像入力部、２は入力した画像を処理して人物の位置を特定する人物位置検出部、３は複数のマイクロホンからの音声を並列に入力する音声入力部、４は画像処理により検出された人物位置に基づいて複数のマイクロホンからの入力音声から雑音を抑圧して音声を取り出す雑音抑圧部である。この雑音抑圧部４の内部構成は、人物位置検出部２により特定された人物位置の中から処理対象とする人物位置を選択する人物位置選択部４－１と、選択された人物位置に基づいて適応フィルタの拘束条件の設定を行う拘束条件設定部４－２と、設定された拘束条件のもとで適応フィルタにより雑音抑圧処理を行う雑音除去部４－＊３とからなる。

【０１３３】人物位置選択部４－１では、第１実施形態において述べたように、画像から得られた複数の人物位置から定めた数の人物位置の選択を行い、拘束条件設定部４－２ではこの人物位置に基づき、線形拘束条件付き適応フィルタの拘束条件を設定する。拘束条件によって、任意の人物位置から到来する音波に対する感度を設定できるようになる。雑音除去部４－３では、設定された拘束条件のもとで適応フィルタにより雑音抑圧処理を行う。

【０１３４】線形拘束条件付き適応フィルタの詳細は、例えば、文献（Heykin著：AdaptiveFilter Theory）に詳述されているが、一応、処理方法を述べる。

【０１３５】式（１）から（４）を参照し、マイクロホンアレイの出力をX、フィルタ係数をW、フィルタの出力をy＝W^H Xとすると、拘束条件付き最小分散適応フィルタのフィルタ係数は、次式によるフィルタの出力パワーy^2 の期待値

$$E[y^2]=E[w^H XX^H w]=w^H Rw \quad (E[\ ]\text{は期待値}) \tag{26}$$

※で、R＝E［XX^H］はXの自己相関行列である。また、拘束条件は、

$$W^H A=g \tag{26-1}$$

と表される。ここで、gは拘束条件の数Gの大きさの定数値の列ベクトルで、例えば［1，1，…，1］であ★り、Aは異なった周波数に関する方向制御ベクトルa_mを列ベクトルとする行列であり、

$$A=[a_1, \cdots, a_L] \tag{27}$$

と表される。上式（６）の成分の各方向制御ベクトルa_m（m＝１，…，L）は

【数１】

$$a_m = (1,\ a_2 e^{-j\omega_m \tau_2},\ \ldots,\ a_N e^{-j\omega_m \tau_N}) \tag{28}$$

【０１３６】である。ここで、τ_2，…，τ_N は１番目のマイクロホンを基準としたときの各マイクロホンに入射する音波の伝搬時間差、ω_m は角周波数、a_2，…，a_N は１番目のマイクロホンを基準としたときの各マイクロホンに入射する音波の振幅比である。Gは例えば１０を用い、ω_m は例えばω_m＝（（ω_a－ω_b）／（G☆－１））＊m＋ω_b とする。ここでω_a は帯域の上限、ω_b は下限の角周波数である。

【０１３７】式（２６－１）の拘束条件として、一つの方向または位置から到来する音波に関する応答を一定にするだけでなく、複数の方向または位置から到来する音波に対する応答を同時に一定にするようにする。例えば、θ_1，θ_2 の２つの到来角度に関する時間遅れτ_i（θ_1），τ_i（θ_2）（式（６））を用いた方向制御ベクトルa_m（θ_1），a_m（θ_2）（m＝０，１，…，L）、

【数２】

$$a_m(\theta_1) = (1,\ a_1 e^{-j\omega_m \tau_1(\theta_1)},\ a_2 e^{-j\omega_m \tau_2(\theta_1)},\ \ldots,\ a_{N-1} e^{-j\omega_m \tau_{(N-1)}(\theta_1)}) \tag{29}$$

$$a_m(\theta_2) = (1,\ a_1 e^{-j\omega_m \tau_1(\theta_2)},\ a_2 e^{-j\omega_m \tau_2(\theta_2)},\ \ldots,\ a_{N-1} e^{-j\omega_m \tau_{(N-1)}(\theta_2)}) \tag{30}$$

【０１３８】を用いて、

$$A=[a_0(\theta_1), a_1(\theta_1), \cdots, a_L(\theta_1), a_0(\theta_2), a_1(\theta_2), \cdots, a_L(\theta_2)] \tag{30}$$

とすることで、複数の到来方向に対するアレイの応答の拘束条件を設定することができる。

【０１３９】ここで、式（４）と（５）による最小化問題を反復的に求める場合、ｊ回めの反復による更新後のフィルタ係数は、次式のように表される。

【０１４０】

$$W_j = P\,[W_{j-1} - \mu y_j X] + F \qquad (31)$$

ここで、ＰとＦは、

$$P = I - A\,(A^H A)^{-1} A^H,\quad F = A\,(A^H A)^{-1} g \qquad (32)$$

である。式（８）により、雑音を抑圧して目的の音声を取り出すフィルタ係数が得られるとともに、雑音を抑圧した音声出力y_jが同時に得られることになる。次に、雑音抑圧のための拘束条件の設定について説明する。複数の人物位置が得られた場合の雑音抑圧の仕方は、第１実施形態で述べたように、処理対象として選択されたすべての人物位置からの到来音波を高い感度で得るようにする抑圧処理モードＡと、選択された人物位置の中の一つから到来する音波のみ高い感度にし他の人物位置からの音波に対しては感度を低くする抑圧処理モードＢがある。他にも、ＡとＢの中間の方法として、所定の複数の人物位置に対して感度を高くしその他に対しては低くするなどが考えられるが、ＡとＢの組み合わせで実現できる。

【０１４１】雑音抑圧の拘束条件は、拘束条件を表す式（２５）において、行列Ａの要素と定数ベクトルｇを与えることにより設定する。処理モードＡもＢも、行列Ａの内容は同じであり、選択した人物位置に関する方向制御ベクトル式（３０）である。定数ベクトルｇの内容は抑圧処理モードに応じて変えるようにし、選択した人物位置すべてに対して感度を高くする抑圧モードＡの場合、ｇの要素はすべて１とし、抑圧処理モードＢの場合、高い感度に設定する人物位置に関するｇの要素は１とし、低い感度に設定する人物位置に関するｇの要素は０にする。

【０１４２】例えば、方向θ_1，θ_2に関する方向制御ベクトルの行列Ａが、次に示す式（３０）の内容の場合、

$A = [a_0(\theta_1), a_1(\theta_1), \cdots, a_L(\theta_1), a_0(\theta_2), a_1(\theta_2), \cdots, a_L(\theta_2)]$

方向θ_1に対して感度を高くし、θ_2に対して感度を低くする場合の定数ベクトルｇの内容は、

$g = [1, 1, \cdots, 1, 0, 0, \cdots, 0]$

とする。

【０１４３】次に、以上の処理の流れについて図２０を参照しながら説明する。

【０１４４】第１実施形態で述べたように、画像から人物位置を特定する画像の処理と、音声の雑音抑圧処理とは、並列に処理するようにし、画像処理の方は第１実施形態と同じである。

【０１４５】まず、初期設定を行い、選択する人物位置の数Ｎと雑音抑圧処理のモードＡかＢかを設定する（ステップＳ５１）。

【０１４６】画像の処理の方では、画像データを、例えば毎秒５フレームで取り込み（ステップＳ５２）、フレームごとに人物位置を特定する（ステップＳ５３）。

【０１４７】音声処理の方では、まず、音声データを例えばサンプリング１２ｋＨｚ、１チャネルあたり１０２４サンプルを１ブロックとして１ブロック分取り込む（ステップＳ５４）。

【０１４８】次に、人物位置が特定されているか否かを判定し（ステップＳ５５）、人物位置が特定されていない場合は、ステップＳ５４に戻り、特定されている場合は次のステップＳ５６に進む。

【０１４９】次のステップＳ５６では、人物位置選択部４－１において、処理対象とする人物位置を選択する。次に、選択された人物位置に基づいてフィルタ処理の拘束条件を式（２６）、（３０）に従って設定する（ステップＳ５７）。

【０１５０】次に、ステップＳ５７で設定した拘束条件のもとに適応フィルタの演算を行って音声を出力する（ステップＳ５８）。以上ステップＳ５２からＳ５３の処理とＳ５４からＳ５８までの処理とを並列に繰り返す。

【０１５１】以下に、拘束条件付きと複数のフィルタを備えた第７実施形態を詳細に説明する。第７実施形態は、拘束条件付き適応フィルタを使った場合に、複数の人物位置各々からの到来音を分離して取り出すものである。複数の人物位置からの到来音を分離して取り出すため、全体構成図の雑音除去部４－３を複数の適応フィルタにより図２１に示すように構成する。

【０１５２】図２１において、適応フィルタの数は人物位置選択部４－１において選択する人物位置の数と一致させてＮ個、例えば３とし、適応フィルタごとに異なった内容で拘束条件を設定する。拘束条件は、拘束条件設定部４－２において行い、ｋ番目の適応フィルタには、選択された人物位置の中のｋ番目の位置に対して感度を高くし、他の人物位置に対しては感度を低くするように設定した拘束条件を入力する。

【０１５３】以上のように、複数の適応フィルタを用い、各々に異なった拘束条件を設定することにより、人物位置ごとの到来音を他の位置からの到来音と分離して抽出することが可能となる。

【０１５４】以下に、拘束条件付き適応フィルタとパイロット信号法を組み合わせた第８実施形態を説明する。第８実施形態の構成を図２２に示す。図２２において、１は画像入力部、２は人物位置検出部、３は音声入力部、４は雑音抑圧部であり、雑音抑制部４の内部は、人物位置選択部４－１と、環境適応部４－２と、雑音除去部４－３と、拘束条件設定部４－５とからなる。

【０１５５】この構成において、画像から検出して人物位置の中から人物位置選択部４－１により複数を選択し、これに基づいて拘束条件設定部４－５において線形

拘束条件付き適応フィルタの拘束条件を設定し、環境適応部４－２においてこの適応フィルタの学習信号である入力信号と希望応答を生成して適応フィルタに入力して雑音抑圧のためのフィルタ係数を決定し、決定したフィルタ係数を雑音除去部４－３に送り、入力音声をフィルタ処理して雑音を除去する。

【０１５６】拘束条件と学習信号の作成方法は種々考えられるが、ここでは、抑圧処理のモードＢの場合について説明する。この場合、一つの人物位置に対して感度を高くなるように拘束条件を設定し、他の人物位置に対しては感度が低くなるように学習信号を生成して適応フィルタの係数を決定する。

【０１５７】例えば、方向θ_1に関して感度を高く設定する場合、方向制御ベクトルの行列Ａを、次に示す内容にし、

$A=[a_0(\theta_1), a_1(\theta_1), \cdots, a_L(\theta_1)]$

定数ベクトルｇの内容は、

$g=[1, 1, \cdots, 1]$

とする。

【０１５８】また、学習信号のうち、入力信号は、感度を低く設定する人物位置からのみの音波の到来を仮定し、マイクロホン位置で観測される信号を第１実施形態の式（１３）により求める。この場合、希望応答は使わないので生成しない。従って、環境適応部は、第１、第２実施形態の環境適応部における希望応答生成部を除いた図２３に示すような構成により実現可能である。

【０１５９】図２３において、４－２ｃは学習信号レベル計算部、４－２ａは入力信号生成部、４－２ｂは音源信号生成部、４－２ｅは適応処理部である。この構成により、拘束条件付き適応フィルタの拘束条件を設定した後、入力信号を適応フィルタに入力してフィルタ係数の更新を行う。

【０１６０】上述の環境適応部４－２の処理の流れを図２４を参照しながら説明する。

【０１６１】まず、環境適応部４－２の音源信号生成部４－２ｂにおいて、選択人物位置の数の系列の音源信号を生成する（ステップＳ６１）。

【０１６２】次に、学習信号レベル計算部４－２ｃにおいて、複数チャネルで入力した入力音声のパワーを計算し、音声検出を行って音声部分と非音声部分とを決め、非音声部分の平均パワーから背景雑音パワーP_Nを求める。このとき、式（１２）により平均化してもよい（ステップＳ６２）。

【０１６３】次に、学習信号レベル計算部４－２ｃにおいて、式（１１）により音源信号の振幅A_kを計算する（ステップＳ６３）。次に、入力信号生成部４－２ａにおいて、式（１３）により、実際の入力音声と加算して適応フィルタの入力信号を生成する（ステップＳ６４）。次に、生成した入力信号を適応制御信号として適応フィルタに入力し、フィルタ係数を得る（ステップＳ６５）。

【０１６４】得られたフィルタは、雑音除去部４－３に送り、入力音声とフィルタを畳み込んで音声を抽出する。雑音除去部４－３におけるフィルタ処理は、式（４）に従って行う。

【０１６５】次に、本実施例全体の処理の流れについて、図２５を参照して説明する。

【０１６６】第１実施形態で述べたように、画像から人物位置を特定する画像の処理と、音声の雑音抑圧処理とは、並列に処理するようにし、画像処理の方は第１実施形態と同じである。

【０１６７】まず、初期設定を行い、選択する人物位置の数Ｎと雑音抑圧処理のモードＡかＢかを設定する（ステップＳ７１）。

【０１６８】画像の処理の方では、画像データを、例えば毎秒５フレームで取り込み（ステップＳ７２）、フレームごとに人物位置を特定する（ステップＳ７３）。

【０１６９】音声処理の方では、まず、音声データを例えばサンプリング１２ｋＨｚ、１チャネルあたり１０２４サンプルを１ブロックとして１ブロック分取り込む（ステップＳ７４）。

【０１７０】次に、人物位置が特定されているか否かを判定し（ステップＳ７５）、人物位置が特定されていない場合は、ステップＳ７４に戻り、特定されている場合は次のステップＳ７６に進む。

【０１７１】次のステップＳ７６では、人物位置選択部において、処理対象とする人物位置を選択する。

【０１７２】次に、選択された人物位置に基づいてフィルタ処理の拘束条件を式（２６）、（３０）に従って設定する（ステップＳ７７）。

【０１７３】次に、環境適応部４－２において、人物位置選択部４－１で選択された人物位置または方向と距離を用いて適応フィルタの学習信号を生成する（ステップＳ７８）。

【０１７４】ステップＳ７７で設定した拘束条件のもとに適応フィルタの演算を行ってフィルタ係数を更新し、雑音除去部４－３にフィルタ係数を転送する（ステップＳ７９）。

【０１７５】次に、雑音除去部４－３において、ステップＳ７９で転送されたフィルタと入力音声との畳み込み演算を行って音声を出力する（ステップＳ８０）。

【０１７６】以上のステップＳ７２からＳ７３の処理と、Ｓ７４からＳ８０までの処理を並列に繰り返す。

【０１７７】以上に述べたように、拘束条件付き適応フィルタにパイロット信号法を適用することにより、パイロット信号法で必要な学習信号生成のための処理量を減らし、また、拘束条件付き適応フィルタにおいて、拘束条件が多い場合のフィルタの自由度低下による性能低下を避けることができるため、少ない処理量で精度よく実

環境に適応して雑音抑圧処理を行うことができる。

【０１７８】以下に第９実施形態を説明する。第９実施形態は、拘束条件付き適応フィルタとパイロット信号法を組み合わせた場合に、複数の人物位置からの音声を分離して取り出すものである。複数の人物位置からの音声を分離して取り出すため、雑音除去部４－３と環境適応部４－２の適応処理部４－２ｅとを、図２６に示すように複数のフィルタから構成するようにしている。

【０１７９】図２６において、適応処理部４－２ｅの適応フィルタと雑音除去部４－２のフィルタはＮ個、例えば３個づつ用意し、適応処理部４－２ｅで決定した適応フィルタの係数を雑音除去部４－３に送るようにしている。適応処理部４－２ｅの各適応フィルタの入力には、入力信号生成部４－２ａで生成した入力信号を共通に入力し、拘束条件は適応フィルタごとに異なったものを入力する。

【０１８０】拘束条件は拘束条件設定部４－５において設定し、ｋ番目の適応フィルタには、人物位置選択部４－１において選択されたｋ番目の人物位置から到来する音波に対して感度を高くなるようにした拘束条件が入力される。拘束条件の設定方法は第８実施形態で述べたのと同じであり、また、適応フィルタの入力信号の生成方法も同じである。

【０１８１】以上に述べたように、拘束条件付き適応フィルタとパイロット信号法を組み合わせた場合に、複数のフィルタにより適応処理を行っているため、人物位置ごとの到来音を分離して抽出することができ、且つ、パイロット信号法で必要な学習信号生成のための処理量を減らし、また、拘束条件付き適応フィルタにおいて、拘束条件が多い場合のフィルタの自由度低下による性能低下を避けることができるため、少ない処理量で精度よく実環境に適応して雑音抑圧処理を行うことができる。

【０１８２】なお、上記した人物位置決定工程と、人物位置選択工程と、フィルタ係数決定工程と、音声抽出工程とはコンピュータプログラムとして、ハードディスク、フロッピーディスク、ＣＤ－ＲＯＭなどの記憶媒体に記憶し、この記憶媒体を適当な計算機に搭載して実行することができる。

【０１８３】

【発明の効果】本発明によれば、複数の人物位置からの音声に対して、背景雑音を抑えてすべての音声を同時に抽出するかあるいは、特定の人物位置からの音声のみを抽出することができる。

【図面の簡単な説明】

【図１】本発明の第１実施形態に係る音声収集装置の構成を示す図である。

【図２】カメラとマイクの配置を示す図である。

【図３】画面上の人物データの一例を示す図である。

【図４】人物位置データの一例を示す図である。

【図５】環境適応部の構成を示す図である。

【図６】フィルタの構成を示す図である。

【図７】マイクロホンと人物位置の設定を示す図である。

【図８】環境適応部の処理の流れを示すフローチャートである。

【図９】第１実施形態全体の処理の流れを示すフローチャートである。

【図１０】第２実施形態における雑音抑圧部の構成を示す図である。

【図１１】環境適応部の構成を示す図である。

【図１２】位置別到来パワーに基づく学習信号の生成について説明するための図である。

【図１３】環境適応部の処理の流れを示すフローチャートである。

【図１４】第３実施形態の全体構成を示す図である。

【図１５】環境適応部の処理の流れを示すフローチャートである。

【図１６】第４実施形態の全体構成を示す図である。

【図１７】第４実施形態の全体処理の流れを示すフローチャートである。

【図１８】第５実施形態における雑音除去部と環境適応部の構成を示す図である。

【図１９】第６実施形態の全体構成を示す図である。

【図２０】第６実施形態の処理の流れを示すフローチャートである。

【図２１】第７実施形態の雑音除去部の構成を示す図である。

【図２２】第８実施形態の全体構成を示す図である。

【図２３】第８実施形態における環境適応部の構成を示す図である。

【図２４】環境適応部の処理の流れを示すフローチャートである。

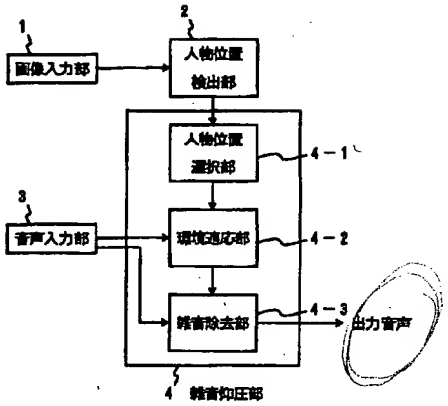
【図２５】第８実施形態の処理の流れを示すフローチャートである。

【図２６】第９実施形態における雑音除去部と環境適応部の構成を示す図である。

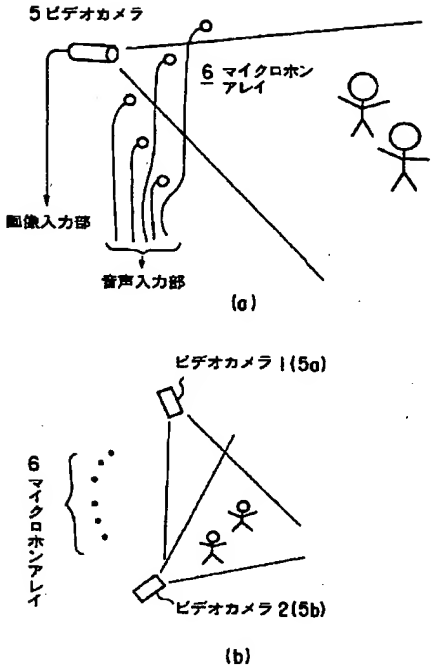
【符号の説明】

１…画像入力部、２…人物位置検出部、３…音声入力部、４…雑音抑圧部、４－１…人物位置検出部、４－２…環境適応部、４－３…雑音除去部。

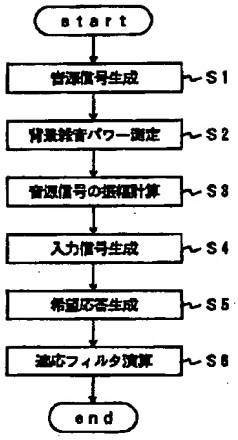
【図１】


1
画像入力部
2
人物位置
検出部
人物位置
選択部
4－1
3
音声入力部
環境適応部
4－2
雑音除去部
4－3
出力音声
4
雑音抑圧部

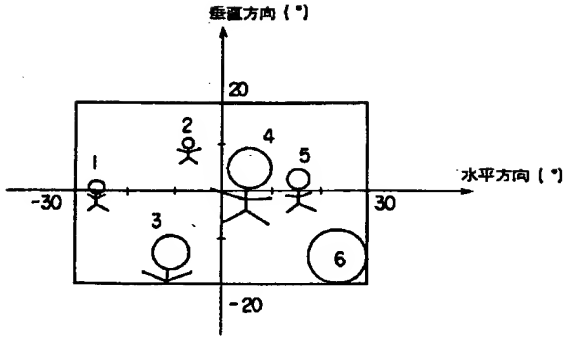
【図２】


5 ビデオカメラ
6 マイクロホン
アレイ
画像入力部
音声入力部
(a)
ビデオカメラ１(5a)
6 マイクロホンアレイ
ビデオカメラ２(5b)
(b)

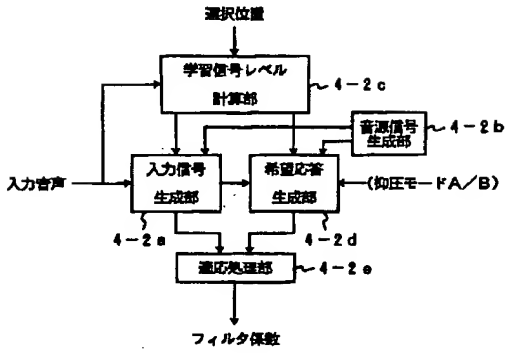
【図８】


start
音源信号生成
S1
背景雑音パワー測定
S2
音源信号の振幅計算
S3
入力信号生成
S4
希望応答生成
S5
適応フィルタ演算
S6
end

【図３】


垂直方向（°）
20
2
4
1
5
水平方向（°）
-30
30
3
6
-20

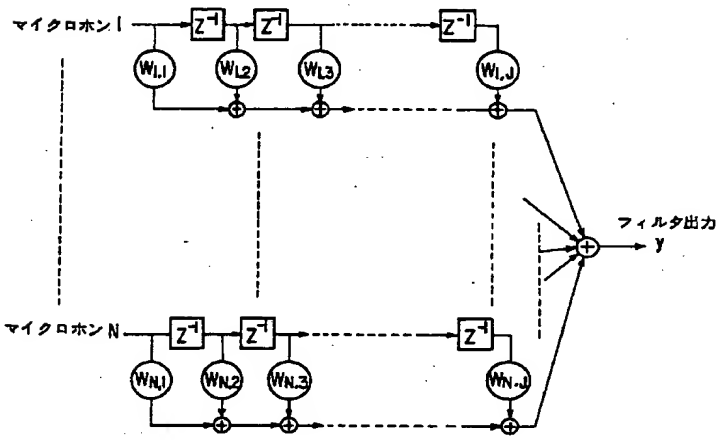
【図５】


選択位置
学習信号レベル
計算部
4－2c
音源信号
生成部
4－2b
入力音声
入力信号
生成部
希望応答
生成部
(抑圧モードA／B)
4－2a
4－2d
適応処理部
4－2e
フィルタ係数

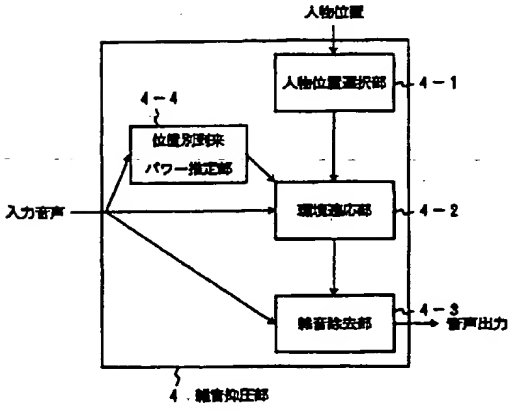
【図４】

番号	人物方向 水平（X）	人物方向 垂直（Y）	A 人物の顔部分の面積（画素数）	B カメラ中心軸方向と人物方向差 $(\sqrt{X^2+Y^2})$	A／B
1	−25°	0°	80	25	3.2
2	− 7°	10°	65	12.2	5.3
3	−10°	−12°	300	15.6	19.2
4	5°	5°	400	7.1	56.3
5	15°	3°	120	15.3	7.8
6	25°	−15°	500	29.2	17.1

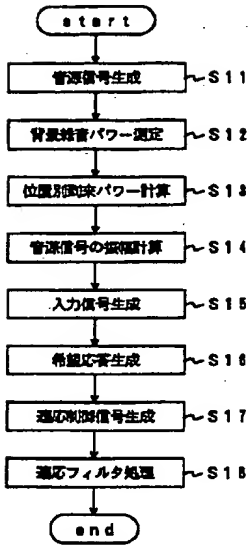
【図６】



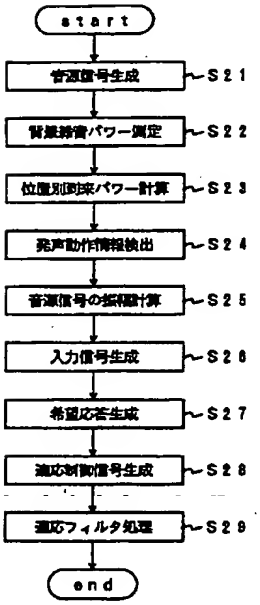
【図１０】



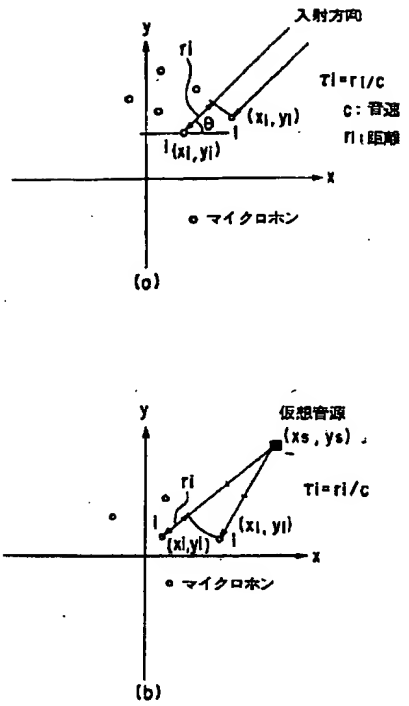
【図１３】



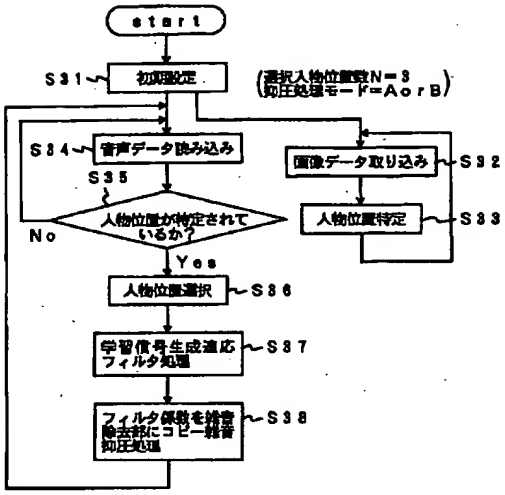
【図１５】



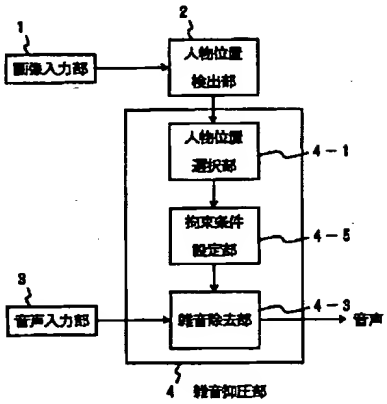
【図７】


入射方向
y
ri
τi=ri/c
c：音速
ri：距離
θ
(xi, yi)
(xi, yi)
x
マイクロホン
(a)
y
仮想音源
(xs, ys)
ri
τi=ri/c
(xi, yi)
(xi, yi)
x
マイクロホン
(b)

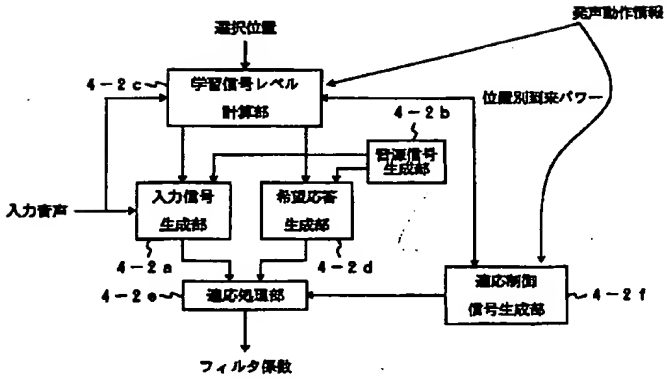
【図９】


start
S31
初期設定
(選択入物位置数N＝3
抑圧処理モードF=AorB)
S34
音声データ読み込み
S35
人物位置が特定されて
いるか？
No
Yes
画像データ取り込み
S32
人物位置特定
S33
人物位置選択
S36
学習信号生成適応
フィルタ処理
S37
フィルタ係数を雑音
除去部にコピー雑音
抑圧処理
S38

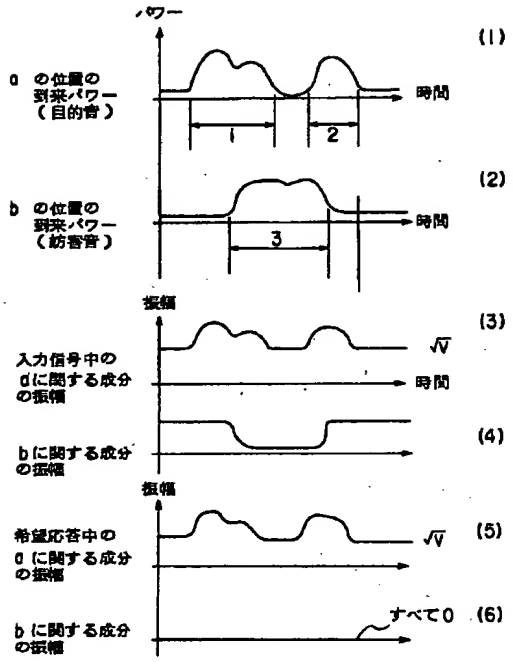
【図１９】


1
画像入力部
2
人物位置
検出部
人物位置
選択部
4－1
拘束条件
設定部
4－5
3
音声入力部
雑音除去部
4－3
音声
4
雑音抑圧部

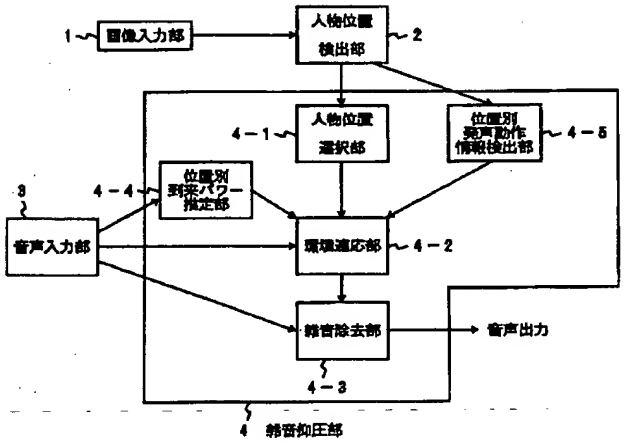
【図１１】


選択位置
発声動作情報
4－2c
学習信号レベル
計算部
位置別到来パワー
4－2b
雑音信号
生成部
入力音声
入力信号
生成部
希望応答
生成部
4－2a
4－2d
4－2e
適応処理部
適応制御
信号生成部
4－2f
フィルタ係数

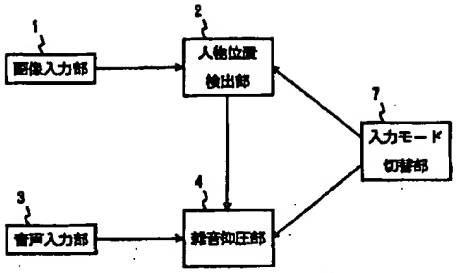
【図１２】



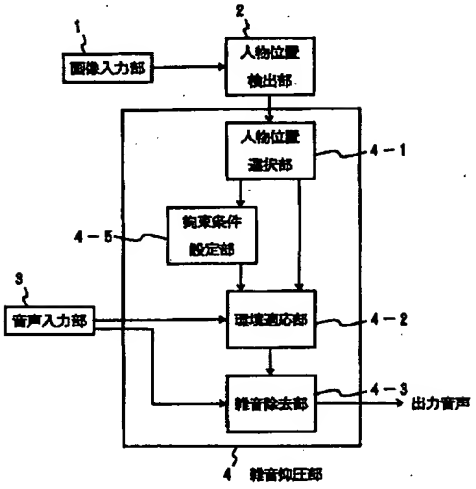
【図１４】



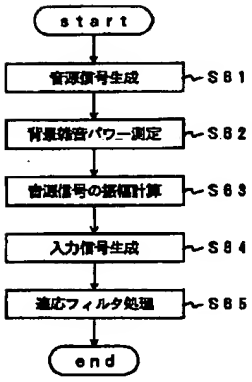
【図１６】



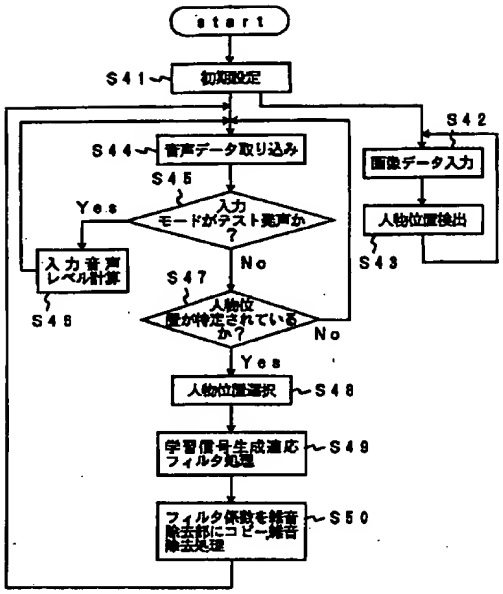
【図２２】



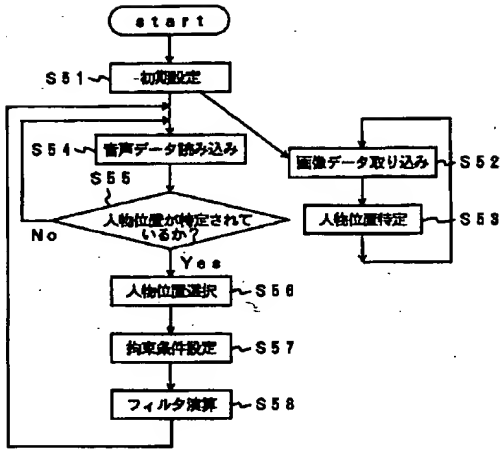
【図２４】



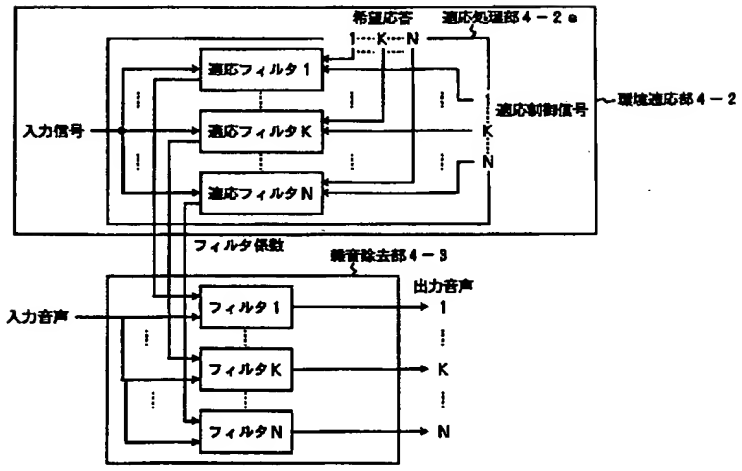
【図１７】


start
S41
初期設定
S42
S44
音声データ取り込み
画像データ入力
S45
Yes
入力モードがテスト発声か？
人物位置検出
入力音声レベル計算
No
S43
S47
人物位置が特定されているか？
S46
No
Yes
人物位置選択
S48
学習信号生成適応フィルタ処理
S49
フィルタ係数を雑音除去部にコピー雑音除去処理
S50

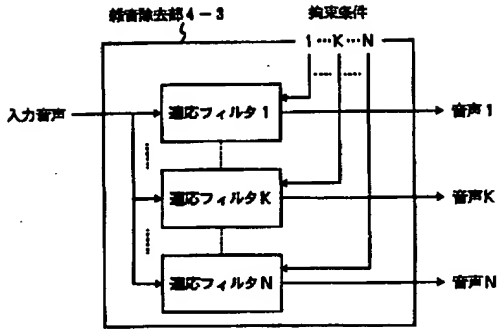
【図２０】


start
S51
初期設定
S54
音声データ読み込み
画像データ取り込み
S52
S55
人物位置が特定されているか？
人物位置特定
S53
No
Yes
人物位置選択
S56
拘束条件設定
S57
フィルタ演算
S58

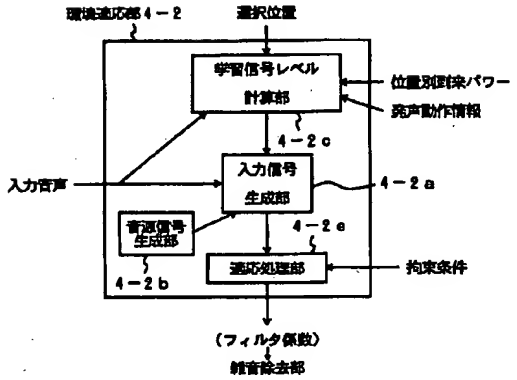
【図１８】


希望応答
適応処理部４－２ａ
1…K…N
適応フィルタ１
適応フィルタＫ
適応フィルタＮ
入力信号
適応制御信号
環境適応部４－２
フィルタ係数
雑音除去部４－３
出力音声
入力音声
フィルタ１
フィルタＫ
フィルタＮ
1
K
N

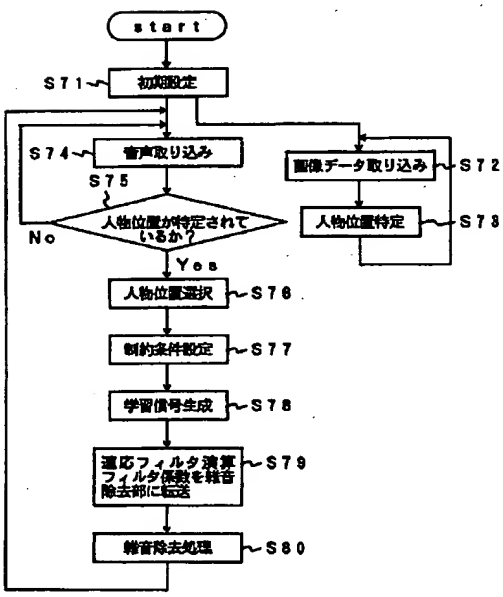
【図２１】


雑音除去部４－３
拘束条件
1…K…N
入力音声
適応フィルタ１
適応フィルタK
適応フィルタN
音声１
音声K
音声N

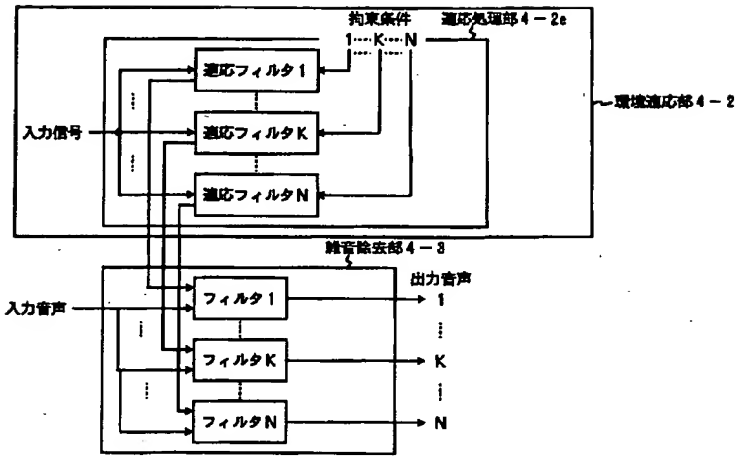
【図２３】


環境適応部４－２
選択位置
学習信号レベル
計算部
位置別雑音パワー
発声動作情報
４－２ｃ
入力音声
入力信号
生成部
４－２ａ
雑音信号
生成部
４－２ｅ
４－２ｂ
適応処理部
拘束条件
（フィルタ係数）
雑音除去部

【図２５】


start
Ｓ７１
初期設定
Ｓ７４
音声取り込み
Ｓ７５
人物位置が特定されて
いるか？
No
Yes
画像データ取り込み
Ｓ７２
人物位置特定
Ｓ７３
人物位置選択
Ｓ７６
制約条件設定
Ｓ７７
学習信号生成
Ｓ７８
適応フィルタ演算
フィルタ係数を雑音
除去部に転送
Ｓ７９
雑音除去処理
Ｓ８０

【図２６】


拘束条件
適応処理部４－２ｅ
1 K N
適応フィルタ１
適応フィルタＫ
適応フィルタＮ
入力信号
環境適応部４－２
雑音除去部４－３
出力音声
入力音声
フィルタ１
フィルタＫ
フィルタＮ
1
K
N